



Safe Offline Reinforcement Learning using Behavior Regularisation and Latent Feasibility Guidance

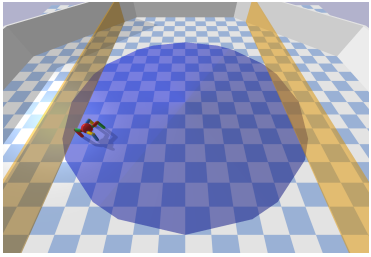
Mahesh Keswani Parth Bhardwaj Ankur Kumar Raunak P. Bhattacharyya

Yardi School of Artificial Intelligence & Dept. of Mathematics and Computing
Indian Institute of Technology Delhi

NeurIPS 2026

Safe offline RL: learning safe policies from fixed datasets

- Robotics, autonomous driving, and industrial control require policies that satisfy explicit safety constraints.
- **Safe offline RL** learns from a fixed dataset of safe and unsafe behavior, without environment interaction.



A safe offline RL method must

1. improve the task return,
2. satisfy the safety constraints,
3. remain close to the data distribution.

Key question

How can a policy be extracted from learned reward and safety critics so that it satisfies the constraints and limits the effect of critic estimation errors?

Problem setting: safe offline RL in a constrained MDP

Constrained MDP

$\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, c, h, \gamma, \rho_0)$

- $\mathcal{S}, \mathcal{A}$: state and action spaces
- $P(s' \mid s, a)$: transition dynamics
- $r(s, a)$: reward, $c : \mathcal{S} \to \mathbb{R}_{\geq 0}$: cost
- $h : \mathcal{S} \to \mathbb{R}$: constraint violation function, $h(s) > 0$ denotes a violation
- $\gamma \in (0, 1)$: discount, ρ_0: initial state distribution

Offline setting. A static dataset $\mathcal{D} = \{(s, a, s', r, c)\}$ collected by an unknown behavior policy π_β, with safe and unsafe trajectories. No environment interaction during training.

Reward and cost value functions

$$V_r^\pi(s) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s\right],$$

$$V_c^\pi(s) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t c(s_t) \,\middle|\, s_0 = s\right].$$

Safe offline RL objective

$$\max_\pi \underbrace{\mathbb{E}_{s\sim\rho_0}\left[V_r^\pi(s)\right]}_{\text{return}} \quad \text{s.t.} \quad \underbrace{\mathbb{E}_{s\sim\rho_0}\left[V_c^\pi(s)\right] \leq \kappa}_{\text{cost limit}}, \qquad \underbrace{D(\pi \,\|\, \pi_\beta) \leq \epsilon}_{\text{behavior regularization}}$$

The cost limit constrains the expected cost. The next slide characterizes safety at the level of individual states.

State-wise feasibility identification [Zheng et al., ICLR 2024]

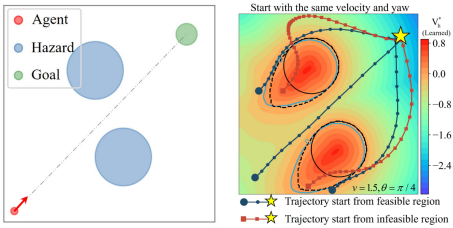
Constraint violation function from the per-step cost:

$$h(s) = \begin{cases} -1, & c(s) = 0, \\ M, & c(s) > 0. \end{cases}$$

Hamilton–Jacobi reachability: worst-case future violation, minimized over policies,

$$V_h^*(s) = \min_{\pi} \max_{t \geq 0} h(s_t), \qquad V_h^*(s) = \min_{a} Q_h^*(s, a).$$

Largest feasible set: $\mathcal{S}_f^* = \{\, s : V_h^*(s) \leq 0 \,\}$, the states from which some policy satisfies the constraint at every future step. The **hazard critics** $\hat{V}_h, \hat{Q}_h$ are learned estimates of V_h^*, Q_h^*.



Right: learned V_h^* for a fixed initial velocity and heading. Trajectories starting in the feasible region avoid the hazards.

Feasible states ($\hat{V}_h(s) \leq 0$)

Maximize return while remaining feasible:

$$\max_{\pi} \mathbb{E}_{a \sim \pi}\left[\hat{Q}_r(s, a)\right] \quad \text{s.t.} \quad \hat{Q}_h(s, a) \leq 0$$

Infeasible states ($\hat{V}_h(s) > 0$)

Minimize future constraint violation:

$$\min_{\pi} \mathbb{E}_{a \sim \pi}\left[\hat{Q}_h(s, a)\right]$$

The policy extraction step affects how well critics are used

Advantage-weighted regression (AWR) [Peng et al., 2019]

$$\max_{\pi} \; \mathbb{E}_{(s,a)\sim\mathcal{D}}\Big[\exp\big(\hat{A}(s,a)/\tau\big)\,\log\pi(a \mid s)\Big]$$

- Weighted behavior cloning of dataset actions.
- The error of $\hat{A}$ enters through the **exponential** weight.

Behavior-regularized policy gradient [Fujimoto & Gu, 2021]

$$\max_{\pi} \; \mathbb{E}_{(s,a)\sim\mathcal{D}}\Big[\lambda\,\hat{Q}\big(s,\pi(s)\big) - \|\pi(s)-a\|_2^2\Big]$$

- Uses the **action gradient** of the critic, with a reconstruction term towards the data.

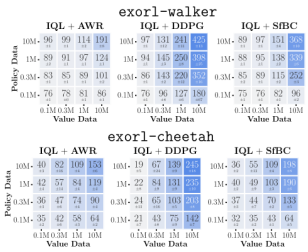


Figure adapted from Park et al., *Is Value Learning Really the Main Bottleneck in Offline RL?*, NeurIPS 2024.

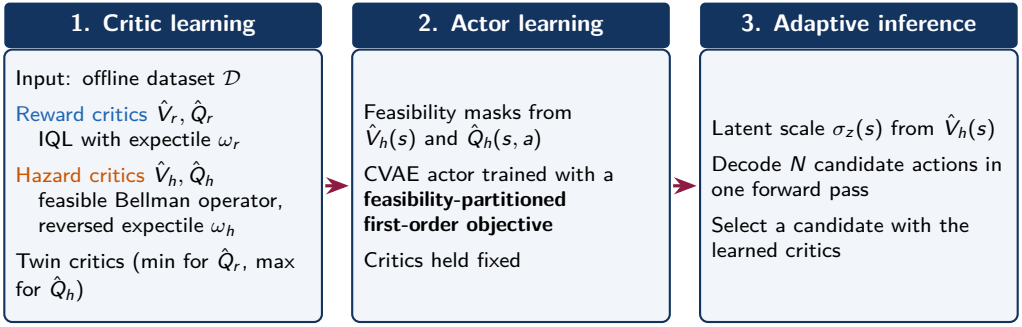
Finding of Park et al. (2024)

The extraction step can affect performance as much as value learning, and behavior-regularized policy gradient often makes more effective use of learned critics.

Policy extraction in safe offline RL poses additional challenges

	Challenge		Approach in FOCAL
1	The policy is guided by a reward critic and a hazard critic acting on the same action. Combining their gradients requires a trade-off weight.	⇒	A **feasibility-partitioned objective**: each transition is guided by exactly one critic, selected by its estimated feasibility.
2	The policy must represent the behavior distribution and support fast inference. Diffusion policies require iterative denoising.	⇒	A **conditional VAE (CVAE) actor** that generates actions in a single forward pass.
3	Offline policies encounter distribution shift at deployment. The degree of conservatism should depend on the state.	⇒	An **adaptive latent prior scale** set by the estimated feasibility of the current state.

FOCAL: Feasibility-Optimized Conditional Actor Learning



Feasibility masks assign each transition to one objective

	$\hat{Q}_h(s,a) \le 0$	$\hat{Q}_h(s,a) > 0$
$\hat{V}_h(s) \le 0$	**Feasible** reward branch	Unsafe action in feasible state *excluded*
$\hat{V}_h(s) > 0$	Contradiction *excluded*	**Infeasible** hazard branch

For a dataset transition (s, a):

$$M_{\text{feas}} = \mathbb{I}\big(\hat{V}_h(s) \le 0 \wedge \hat{Q}_h(s,a) \le 0\big),$$
$$M_{\text{infeas}} = \mathbb{I}\big(\hat{V}_h(s) > 0 \wedge \hat{Q}_h(s,a) > 0\big),$$
$$M_{\text{valid}} = M_{\text{feas}} + M_{\text{infeas}}.$$

- **Contradiction**: $\hat{V}_h(s) > 0$ but $\hat{Q}_h(s,a) \le 0$ is inconsistent with $V_h^*(s) = \min_a Q_h^*(s,a)$.
- **Unsafe action in a feasible state**: not used for policy updates.

A first-order objective partitioned by feasibility

The CVAE actor (encoder q_ϕ, decoder D_θ) reconstructs $\hat{a} = D_\theta(s, z)$, $z \sim q_\phi(\cdot \mid s, a)$, and minimizes

$$\mathcal{L}(\theta, \phi) = \mathbb{E}\Big[M_{\text{feas}}\big(\|a - \hat{a}\|_2^2 - \lambda_r \hat{Q}_r(s, \hat{a})\big) + M_{\text{infeas}}\big(\|a - \hat{a}\|_2^2 + \lambda_h \hat{Q}_h(s, \hat{a})\big) + M_{\text{valid}}\, \beta\, D_{\text{KL}}\big(q_\phi(z \mid s, a) \,\|\, \mathcal{N}(0, I)\big)\Big]$$

with normalized step sizes $\lambda_r = \alpha_r / \Phi_r$, $\lambda_h = \alpha_h / \Phi_h$, where Φ is the mean $|\hat{Q}|$ over the mini-batch (as in TD3+BC).

Feasible transitions

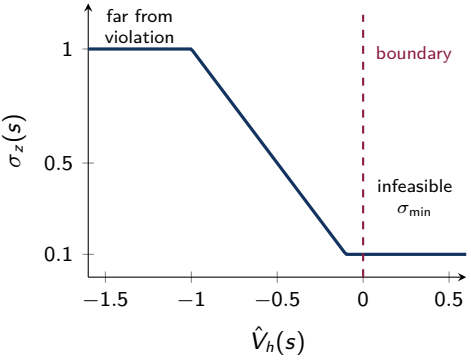
The reward term moves $\hat{a}$ towards higher estimated return. The reconstruction term keeps $\hat{a}$ close to the dataset action, which bounds the hazard value of the training target (Proposition 1).

Infeasible transitions

The hazard term moves $\hat{a}$ towards lower estimated constraint violation, while the reconstruction term limits its deviation from the dataset action.

- Each valid transition receives the action gradient of **exactly one** critic (Proposition 2), so no reward–cost trade-off weight is required.
- The objective contains no exponential weights, so no weight clipping is required.

Adaptive inference: the latent scale follows estimated feasibility



$$\sigma_z(s) = \mathrm{clip}\big(-\hat{V}_h(s),\, \sigma_{\min},\, 1\big)$$

1. Compute $\sigma_z(s)$ from the hazard value function.
2. Sample $z_i \sim \mathcal{N}\big(0, \sigma_z(s)^2 I\big)$, $i = 1, \ldots, N$, and decode $\hat{a}_i = D_\theta(s, z_i)$.
3. Let $C(s) = \{i : \hat{Q}_h(s, \hat{a}_i) \le 0\}$ and execute

$$i^\star = \begin{cases} \arg\max_{i \in C(s)} \hat{Q}_r(s, \hat{a}_i), & C(s) \neq \emptyset, \\ \arg\min_i \hat{Q}_h(s, \hat{a}_i), & C(s) = \emptyset. \end{cases}$$

- Near the boundary, samples concentrate around the mode of the prior.
- All candidates are decoded in one batched forward pass ($N = 16$).
- Filtering reduces, but does not eliminate, constraint violations.

FOCAL algorithm

Input: offline dataset $\mathcal{D}$, gradient steps K, warm-up steps K_{warm}, number of candidates N, minimum latent scale $\sigma_{\min}$

Training

Initialize $\hat{V}_r$, $\hat{V}_h$, twin critics $\{\hat{Q}_{r,i}\}_{i=1,2}$, $\{\hat{Q}_{h,i}\}_{i=1,2}$, target copies, encoder q_ϕ, decoder D_θ
for $k = 1, \ldots, K$ **do**
  Sample a mini-batch $\mathcal{B} \sim \mathcal{D}$
  Update $\hat{V}_r$, $\hat{Q}_{r,i}$ with IQL
  Update $\hat{V}_h$, $\hat{Q}_{h,i}$ with the feasible Bellman operator
  if $k > K_{\text{warm}}$ **then**
    Compute masks $M_{\text{feas}}, M_{\text{infeas}}$ and step sizes λ_r, λ_h on $\mathcal{B}$
    Update ϕ, θ by minimizing $\mathcal{L}(\theta, \phi)$, critics held fixed
  end if
  Update target critics by Polyak averaging
end for

Inference at state s

$\sigma_z(s) \leftarrow \mathrm{clip}\big(-\hat{V}_h(s), \sigma_{\min}, 1\big)$
Sample $z_i \sim \mathcal{N}\big(0, \sigma_z(s)^2 I\big)$, decode $\hat{a}_i = D_\theta(s, z_i)$, $i = 1, \ldots, N$
$C(s) \leftarrow \{\, i : \hat{Q}_h(s, \hat{a}_i) \le 0 \,\}$
if $C(s) \neq \emptyset$ **then**
  Execute $\hat{a}_{i^\star}$, $i^\star = \arg\max_{i \in C(s)} \hat{Q}_r(s, \hat{a}_i)$
else
  Execute $\hat{a}_{i^\star}$, $i^\star = \arg\min_i \hat{Q}_h(s, \hat{a}_i)$
end if

Theory: how critic errors enter the policy update

For a dataset transition (s, a), we compare the update under the learned critic with the update under the exact critic, over critic training randomness.

AWR (unclipped weight)

With $\hat{A} = A^* + \epsilon_A$, $\epsilon_A \sim \mathcal{N}(\mu_A, \sigma_A^2)$, the variance of the update is

$$\|v\|_2^2 \, e^{2A^*/\tau} \, e^{2\mu_A/\tau + \sigma_A^2/\tau^2} \left(e^{\sigma_A^2/\tau^2} - 1 \right).$$

It depends **exponentially** on the advantage error.

FOCAL (first-order objective)

With $\hat{Q} = Q^* + \epsilon_Q$ and error variance $\sigma^2(s, \hat{a})$, the variance of the update satisfies

$$\mathbb{E}_\epsilon \left\| g - \mathbb{E}_\epsilon[g] \right\|_2^2 \leq \lambda^2 \operatorname{tr} F(\psi) \, \mathbb{E}_{\hat{a}} \left[\sigma^2(s, \hat{a}) \right].$$

It is bounded **linearly** in the variance of the critic error.

Action deviation (Proposition 1)

For a feasible transition, the training target $\hat{a}^*$ satisfies $\|\hat{a}^* - a\|_2 \leq \lambda_r L_{Q_r}/2$ and $\hat{Q}_h(s, \hat{a}^*) \leq \hat{Q}_h(s, a) + \lambda_r L_{Q_r} L_{Q_h}/2$.

$F(\psi)$: Fisher information of the actor, with $\operatorname{tr} F \leq d L_\theta^2 / \sigma_{\text{dec}}^2 + \operatorname{tr} F_q$.

FOCAL satisfies the cost limit on all 26 DSRL tasks

Normalized return (↑) and cost (↓), averaged over the tasks of each suite, 3 training runs per task. Safe: cost ≤ 1.

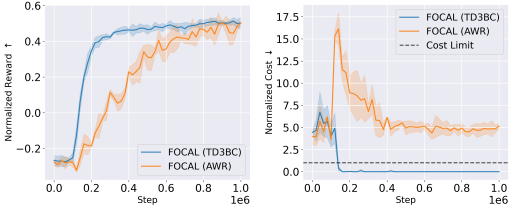
	BC		CDT		BCQ-Lag		CPQ		COptiDICE		LSPC		FISOR		**FOCAL**	
Suite	R	C	R	C	R	C	R	C	R	C	R	C	R	C	R	C
Safety-Gym (9)	0.49	7.67	0.61	3.65	0.52	12.58	0.05	11.51	0.32	6.96	0.49	1.29	**0.33**	**0.39**	**0.47**	**0.39**
Bullet-Gym (8)	0.72	10.58	0.63	2.09	0.71	10.63	0.32	5.28	0.55	7.64	**0.43**	**0.32**	**0.39**	**0.10**	**0.54**	**0.17**
MetaDrive (9)	0.40	3.21	0.25	1.45	0.59	7.48	**-0.06**	**0.24**	0.63	10.26	0.59	1.22	**0.36**	**0.25**	**0.35**	**0.70**

Gray: unsafe on average. Bold: safe. Blue: highest return among safe methods.

- FOCAL and FISOR satisfy the cost limit on every task.
- FOCAL obtains the highest average return among safe methods on Safety-Gymnasium and Bullet-Safety-Gym, and a return comparable to FISOR on MetaDrive with a higher cost.

Ablations: extraction objective and latent scale

First-order objective vs. AWR



Only the extraction objective is replaced. The AWR variant shows larger variation in cost and exceeds the threshold at several points during training. The first-order variant remains below it.

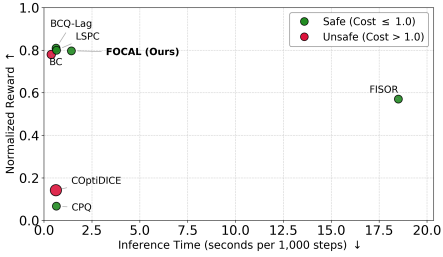
Fixed vs. adaptive latent scale ($N = 16$)

	σ_z=1.0		σ_z=0.5		σ_z=0.1		$\sigma_z(s)$	
Task	R	C	R	C	R	C	R	C
AntVel	0.76	0.00	0.77	0.00	0.74	0.00	**0.81**	0.00
CarGoal1	0.62	1.63	0.62	1.36	0.59	1.43	**0.61**	0.80
CarRun	0.82	0.00	0.82	0.00	0.75	0.00	**0.83**	0.00
DroneRun	0.56	0.80	0.56	1.26	0.58	1.33	**0.58**	0.58

The adaptive scale satisfies the threshold on all four tasks, with a return equal to or higher than every fixed scale that satisfies the threshold.

Single-step inference and further analyses

Inference latency



Time for 1,000 steps on Walker2d velocity, including environment interaction.

Further analyses (appendix)

- **Number of candidates** N: all values in $\{1, 4, 16, 64\}$ satisfy the threshold on five tasks. $N = 16$ is used throughout.
- **Hazard critic accuracy**: a non-zero fraction of unsafe steps is classified as feasible.
- **Dataset size**: FOCAL satisfies the threshold with 25% and 50% of the trajectories, with lower return.
- **Limited boundary coverage**: training on the 20% lowest-cost trajectories satisfies the threshold, with lower return.

Conclusion

Summary

- A **feasibility-partitioned first-order objective** for a single-step CVAE actor.
- An **adaptive latent scale** based on the estimated feasibility of the state.
- Critic errors enter the update **linearly** for FOCAL and **exponentially** for unclipped AWR.
- On DSRL, FOCAL satisfies the cost limit on all evaluated tasks with competitive or higher returns.

Limitations and future work

- Safety depends on the accuracy of the learned feasibility critics, and no state-wise guarantee is provided at deployment.
- Future work: uncertainty estimates for the hazard critic and offline-to-online fine-tuning.

Code

https://github.com/mahesh-keswani/FOCAL-Offline-SafeRL

Thank you!

References

S. Fujimoto and S. S. Gu. A minimalist approach to offline reinforcement learning. NeurIPS, 2021.

S. Fujimoto, D. Meger, and D. Precup. Off-policy deep reinforcement learning without exploration. ICML, 2019.

I. Kostrikov, A. Nair, and S. Levine. Offline reinforcement learning with implicit Q-learning. ICLR, 2022.

P. Koirala, Z. Jiang, S. Sarkar, and C. Fleming. Latent safety-constrained policy approach for safe offline reinforcement learning. ICLR, 2025.

Z. Liu et al. Datasets and benchmarks for offline safe reinforcement learning. DMLR, 2024.

S. Park, K. Frans, S. Levine, and A. Kumar. Is value learning really the main bottleneck in offline RL? NeurIPS, 2024.

X. B. Peng, A. Kumar, G. Zhang, and S. Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. arXiv:1910.00177, 2019.

D. Yu, H. Ma, S. Li, and J. Chen. Reachability constrained reinforcement learning. ICML, 2022.

Y. Zheng, J. Li, D. Yu, Y. Yang, S. Li, X. Zhan, and J. Liu. Safe offline reinforcement learning with feasibility-guided diffusion model. ICLR, 2024.