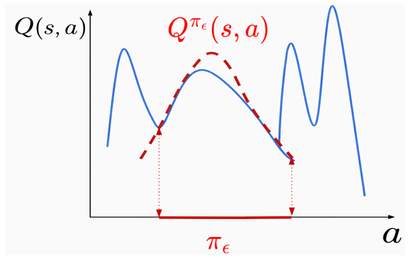
REG: In-Sample RL via Regularizing the Evaluation Gap

Hanpu Shen[1] Weining Shen[1] Roy Fox[2]

[1]Department of Statistics [2]Department of Computer Science UC Irvine

1. Motivating Example

Question. Given a learned Q function, how do we construct a policy π_ϵ that is **plausible under offline data**?



- π_ϵ denotes a candidate policy whose **evaluation error** is at most ϵ.
- The red interval is the action support of π_ϵ; it should stay in regions where Q can be trusted from data.
- The goal is not just high Q value, but small residual throughout the policy support.

Pick a policy with small evaluation error: $J(\pi_\epsilon) - \mathbb{E}_{(s,a)\sim d_0}[Q(s,a)] \le \epsilon.$

$$\underbrace{J(\pi_\epsilon) - \mathbb{E}_{(s,a)\sim d_0}[Q(s,a)]}_{\text{evaluation error}} = \frac{1}{1-\gamma}\underbrace{\mathbb{E}_{(s,a)\sim d_{\pi_\epsilon}}[\Delta^{\pi_\epsilon}Q(s,a)]}_{\text{Bellman residual under } \pi_\epsilon} \le \epsilon.$$

Here $\Delta^{\pi_\epsilon}Q(s,a) = r(s,a) + \gamma\mathbb{E}_{s'\sim P,\, a'\sim\pi_\epsilon}[Q(s',a')] - Q(s,a)$ is the **Bellman error** of Q under π_ϵ. Here d_0 is the initial state-action distribution, and d_π is the discounted occupancy measure of policy π. Since data cannot estimate it perfectly everywhere, REG uses a robust objective in Section 2.

2. Robust Problem Formulation

REG objective. Maximize return while controlling the worst-case evaluation gap over plausible critics. Here f is a convex critic-loss metric, and d_μ is the occupancy measure of the available offline data.

$$\begin{aligned} \max_{\pi\in\Pi} \quad & \mathbb{E}_{d_\pi}[r] \\ \text{s.t.} \quad & \max_{q\in\mathcal{Q}(\pi)} \left|\mathbb{E}_{d_\pi}[\Delta^\pi q]\right| \le \epsilon_1, \\ & \mathcal{Q}(\pi) = \left\{q : \mathbb{E}_{d_\mu}\left[f(\Delta^\pi q)\right] \le \epsilon_2\right\}. \end{aligned}$$

ϵ_1 controls conservativeness; ϵ_2 controls the critic-fit versus uncertainty tradeoff.

Interpretation

- $\mathcal{Q}(\pi)$ contains critics whose Bellman error is plausible on the available offline data.
- The outer constraint rejects policies with large worst-case evaluation gaps.
- This avoids a conservative estimate of supported action values while still guarding against out-of-distribution overestimation.

3. REG Solution and Algorithm

REG solves the robust objective through a **two-stage learning route**: first learn a robust state-value function, then extract a policy from the induced weights. Define the Bellman residual induced by a candidate value function:

$$A_V(s,a) = r(s,a) + \gamma\mathbb{E}_{s'\sim p(\cdot|s,a)}[V(s')] - V(s).$$

Theory details: the full dual form, including the skew direction, is Theorem 4.2 in the paper.

Equivalent square-loss solution

(I) Value learning: $\min_V \mathbb{E}_{d_\mu}\left[\frac{1}{2}\max(0, A_V(s,a))^2\right] + \alpha\,\mathbb{E}_{d_0}[V],$

(II) Target occupancy: $d_{\pi^*}(s,a) \propto d_\mu(s,a)\cdot\max(0, A_V(s,a)).$

Policy extraction with Orthogonal Policy Gradients (OPG).

REG's target occupancy gives a weighted-cloning direction, but Behavior Cloning is mode-covering rather than mode-seeking. OPG combines this stable direction with an orthogonal probing direction:

$$\begin{aligned} \omega(s,a) &= \max(0, Q_{\phi'}(s,a) - V_\psi(s)), \\ \boldsymbol{g}_{\text{wbc}} &= \mathbb{E}_{\mathcal{D}}[\omega(s,a)\nabla_\theta \log\pi_\theta(a|s)], \\ \boldsymbol{g}_{\text{pg}} &= \mathbb{E}_{s\sim\mathcal{D},\, a\sim\pi_\theta}[(Q_\phi(s,a) - V_\psi(s))\nabla_\theta\log\pi_\theta(a|s)], \\ \boldsymbol{g}_{\text{final}} &= \boldsymbol{g}_{\text{wbc}} + \lambda\left(\boldsymbol{g}_{\text{pg}} - \text{proj}_{\boldsymbol{g}_{\text{wbc}}}\boldsymbol{g}_{\text{pg}}\right). \end{aligned}$$

ω: REG-induced cloning weight; $\boldsymbol{g}_{\text{wbc}}$: data-supported weighted cloning; $\boldsymbol{g}_{\text{pg}}$: mode-seeking probe; $\boldsymbol{g}_{\text{final}}$: add only the probe component orthogonal to WBC.

Algorithm: mini-batch REG update with OPG *value → critic → policy*

1 **Initialize** critic Q_ϕ, target critic $Q_{\phi'}$, value V_ψ, and policy π_θ.

2 **Sample** mini-batch $\mathcal{B} = \{(s,a,r,s')\} \sim \mathcal{D}$.

3 **Robust value step.** Update V_ψ by $\psi \leftarrow \psi - \eta_V\nabla_\psi\mathcal{L}_V$:

$$\mathcal{L}_V = \mathbb{E}_{\mathcal{B}}\left[\frac{1}{2}\max(0, Q_{\phi'}(s,a) - V_\psi(s))^2 + \alpha V_\psi(s)\right].$$

Learns the square-loss REG value objective; α controls conservativeness.

4 **Critic step.** Fit Q_ϕ to the value bootstrap and update the target critic:

$$\mathcal{L}_Q = \mathbb{E}_{\mathcal{B}}\left[(r + \gamma V_\psi(s') - Q_\phi(s,a))^2\right], \qquad \phi' \leftarrow \beta\phi + (1-\beta)\phi'.$$

$Q_{\phi'}$ provides a stable estimate of the REG-induced weights.

5 **Policy step.** Apply the conservative direction plus the orthogonal probing component:

$$\theta \leftarrow \theta + \eta_\pi \boldsymbol{g}_{\text{final}}, \qquad \boldsymbol{g}_{\text{final}} = \boldsymbol{g}_{\text{wbc}} + \lambda\boldsymbol{g}_{\text{pg}}^{\perp}.$$

Weighted cloning stays data-supported; OPG uses λ to add mode-seeking improvement without interfering with improvement along the WBC gradient.

4. Suboptimality Bound

Under regularity conditions, the learned policy satisfies the following optimality bound:

$$J_{\pi^*} - J_{\hat{\pi}} \le \frac{1}{1-\gamma}\mathcal{O}\left(\epsilon_1 + \sqrt{\epsilon_2} + \frac{\text{poly}(R,B)}{\sqrt{\epsilon_2}}\left(\mathcal{R}_n(\mathcal{F}_V) + \sqrt{\frac{\log(1/\delta)}{n}}\right)\right).$$

δ: failure prob.; $\mathcal{R}_n(\mathcal{F}_V)$: value-class complexity; poly(R,B): reward/function-class constants.

5. D4RL Results

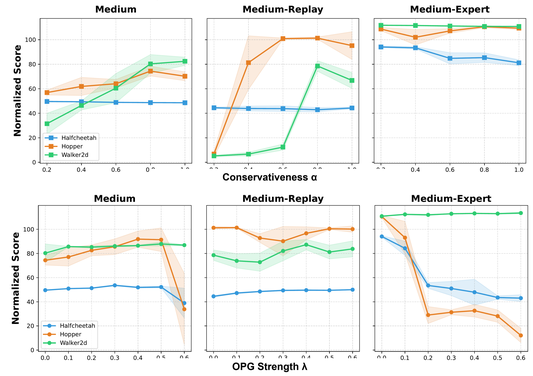
D4RL normalized scores; mean over 5 seeds. Top two: **best**, second best.

	Gaussian Policy (Baselines)					Diffusion Policy		Gaussian Policy
Dataset	**IQL**	**IVR**	**AlignIQL**	**EQL**	**ReBRAC**	**Diffusion-QL**	**IDQL**	**REG (ours)**
halfcheetah-medium	47.4 ± 0.2	48.3 ± 0.2	44.2 ± 0.3	48.3	**65.6 ±1.0**	51.1 ± 0.5	51.0	53.6 ± 0.5
hopper-medium	66.3 ± 5.7	75.5 ± 3.4	57.8 ± 2.4	74.2	**102.0 ±1.0**	90.5 ± 4.6	65.4	91.9 ± 6.7
walker2d-medium	72.5 ± 8.7	84.2 ± 4.6	76.7 ± 3.4	84.2	82.5 ±3.6	87.0 ± 0.9	82.5	**87.9 ± 1.7**
halfcheetah-medium-replay	44.2 ± 1.2	44.8 ± 0.7	37.3 ± 0.2	45.2	**51.2 ±3.2**	47.8 ± 0.3	45.9	50.0 ± 0.2
hopper-medium-replay	95.2 ± 8.6	99.7 ± 3.3	77.9 ± 28.9	100.7	98.1 ±5.3	101.3 ± 0.6	92.1	**101.4 ± 0.5**
walker2d-medium-replay	76.1 ± 7.3	81.2 ± 3.8	66.3 ± 9.1	82.2	77.3 ±7.9	**95.5 ± 1.5**	85.1	87.3 ± 4.3
halfcheetah-medium-expert	86.7 ± 5.3	94.0 ± 0.4	81.9 ± 1.5	94.2	**101.1 ±5.2**	96.8 ± 0.3	95.9	94.1 ± 0.6
hopper-medium-expert	101.5 ± 7.3	**111.8 ± 2.2**	75.2 ± 5.9	111.2	107.0 ±6.4	111.1 ± 1.3	108.6	110.6 ± 0.9
walker2d-medium-expert	110.6 ± 1.0	110.0 ± 0.8	104.4 ± 9.5	112.7	111.6 ±0.3	110.1 ± 0.3	112.7	**113.5 ± 0.8**
Locomotion-Average	77.8	83.3	69.1	83.7	88.4	87.9	81.9	87.8 ± 0.9
antmaze-umaze	77.0 ± 5.5	92.2 ± 1.4	95.6 ± 2.2	93.8	**97.8 ±1.0**	93.4 ± 3.4	94.0	94.1 ± 1.6
antmaze-umaze-diverse	54.3 ± 5.5	74.0 ± 2.3	72.0 ± 7.3	82.0	**88.3 ±13.0**	66.2 ± 8.6	80.2	76.4 ± 6.1
antmaze-medium-play	65.8 ± 11.7	80.2 ± 3.7	**88.0 ± 2.7**	76.0	84.0 ±4.2	76.6 ± 10.8	84.5	85.0 ± 5.2
antmaze-medium-diverse	73.8 ± 5.5	79.1 ± 4.2	83.2 ± 5.2	73.6	76.3 ±13.5	78.6 ± 10.3	**84.8**	84.4 ± 4.9
antmaze-large-play	42.0 ± 4.5	53.2 ± 4.8	55.2 ± 9.5	46.5	60.4 ±26.1	46.4 ± 8.3	63.5	**63.6 ± 8.6**
antmaze-large-diverse	30.3 ± 3.6	52.3 ± 5.2	58.0 ± 3.6	49.0	54.4 ±25.1	56.6 ± 7.6	**67.9**	61.2 ± 5.2
AntMaze-Average	57.2	71.8	75.3	70.2	76.8	69.6	79.2	77.5 ± 2.3

REG is a Gaussian-policy method competitive with diffusion policies, while avoiding iterative diffusion sampling at action selection.

6. Ablations

D4RL MuJoCo locomotion ablations. Medium, Medium-Replay, and Medium-Expert denote dataset quality/mix for HalfCheetah, Hopper, and Walker2d.



- α tunes conservativeness of value learning.
- λ balances behavior-cloning safety with policy-gradient improvement.

7. Takeaway

REG turns a fixed-dataset limitation into a robust optimization principle: **learn values only where Bellman errors are defensible, then improve the policy with a balanced update.**