PAIRED RELATIVE-VALUE REGULARIZATION FOR OFFLINE REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Offline reinforcement learning (RL) suffers from unreliable value estimates for actions that are insufficiently covered by the offline dataset. Conservative value learning mitigates this issue by suppressing the values of such actions, but overly strong regularization can also limit useful policy improvement. In this work, we propose Paired Relative-Value Regularization (PaRE), which approaches conservative value regularization from a policy-local perspective. PaRE constructs candidate actions around the current policy and regularizes their values relative to the data action observed at the same state, allowing the regularization strength to adapt to the local relative-value structure. We theoretically analyze the relationship between PaRE and policy-local value overestimation, and characterize how its gradient determines the overall regularization strength and its allocation across candidates. Experiments on the D4RL MuJoCo locomotion and Adroit dexterous manipulation benchmarks show that PaRE achieves competitive performance across datasets with different data qualities.

1 INTRODUCTION

Offline reinforcement learning (RL) learns policies from previously collected data without further interaction with the environment, making it useful when online data collection is costly or unsafe (Levine et al., 2020). A key challenge is the distribution shift between the learned policy and the behavior policy that generated the data. During policy improvement, the learned policy may favor actions that are poorly represented in the dataset and whose values are overestimated. Without additional interaction to correct these estimates, the errors can propagate through bootstrapped updates and degrade policy performance (Fujimoto et al., 2019; Kumar et al., 2019). Conservative value learning addresses this problem by penalizing potentially overestimated action values (Kumar et al., 2020). However, excessive conservatism can also suppress useful alternatives to the observed actions and limit policy improvement (Lyu et al., 2022; Yao et al., 2025).

Two common approaches to this challenge are policy constraints and value regularization. Policy-constraint methods encourage the learned policy to remain close to actions supported by the offline data (Fujimoto et al., 2019; Kumar et al., 2019; Wu et al., 2019; Fujimoto & Gu, 2021), while conservative value-learning methods, such as CQL, penalize the values of alternative actions relative to observed actions (Kumar et al., 2020). Recent work seeks to reduce excessive conservatism by supporting value generalization through pseudo values in MCQ (Lyu et al., 2022) and local smoothing in SQOG (Yao et al., 2025), or by adapting regularization to state-level data quality in SSAR (Luo et al., 2025). We study this problem from a policy-local perspective, focusing on how conservative regularization should respond to the values of nearby candidate actions relative to the observed action.

In deterministic actor–critic learning, policy updates depend on the critic's action gradient at the current policy action (Silver et al., 2014). We therefore focus on candidate actions in its neighborhood and compare their values with that of the logged action at the same state. These relative-value gaps guide the strength and allocation of conservative regularization.

We propose **Paired Relative-Value Regularization (PaRE)**, a critic-side regularizer based on policy-local value comparisons. PaRE pairs candidate actions around the current policy with the logged action at the same state and applies a smooth penalty to their aggregated value differences.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The regularization places greater weight on candidates with larger relative values, while its overall strength decreases as the aggregated gap becomes smaller. This allows conservative correction to respond to the local relative-value structure without an additional behavior model or uncertainty estimator.

Our contributions are threefold: (1) We introduce a policy-local critic regularizer that pairs candidate actions with logged actions and determines conservative correction from their relative values. (2) We provide a theoretical analysis of PaRE's relationship to policy-local value overestimation and its regularization behavior. (3) We evaluate PaRE on 24 D4RL MuJoCo locomotion and Adroit manipulation tasks spanning different dataset qualities, demonstrating competitive performance across these benchmarks.

2 PRELIMINARIES

Offline reinforcement learning. We consider a discounted Markov decision process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \rho_0, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces, $P(\cdot \mid s, a)$ is the transition kernel, $r(s, a)$ is the reward function, ρ_0 is the initial-state distribution, and $\gamma \in [0, 1)$ is the discount factor. A policy $\pi(a \mid s)$ specifies a distribution over actions given a state. The objective of reinforcement learning is to find a policy that maximizes the expected discounted return

$$J(\pi) = \mathbb{E}_{\pi, P, \rho_0}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right]. \tag{1}$$

The action-value function $Q^\pi(s, a)$ denotes the expected discounted return obtained by taking action a at state s and following π thereafter.

In offline reinforcement learning, the agent learns from a fixed dataset $\mathcal{D} = \{(s_i, a_i, r_i, s'_i, d_i)\}_{i=1}^N$ collected prior to training by one or more behavior policies, where $d_i \in \{0, 1\}$ denotes the terminal indicator. No additional interaction with the environment is allowed during training, and the policy is optimized solely using samples from $\mathcal{D}$. For a sampled transition $(s, a_D, r, s', d) \sim \mathcal{D}$, a_D denotes the logged action at state s. We consider a deterministic policy π_ϕ and write its action as $a_\pi = \pi_\phi(s)$.

Actor–critic framework. Actor–critic methods jointly learn a parameterized policy π_ϕ and an action-value function Q_θ. The critic estimates the value of state–action pairs using temporal-difference learning, while the actor is optimized according to the learned value function. For a deterministic policy, the basic policy improvement objective can be written as

$$\max_\phi \ \mathbb{E}_{s \sim \mathcal{D}}\left[Q_\theta(s, \pi_\phi(s))\right], \tag{2}$$

where $s \sim \mathcal{D}$ denotes sampling states from the empirical state distribution of the offline dataset.

Given a transition $(s, a_D, r, s', d) \sim \mathcal{D}$, the critic is typically trained by minimizing a temporal-difference objective

$$\mathcal{L}_{\mathrm{TD}}(\theta) = \mathbb{E}_{\mathcal{D}}\left[\left(Q_\theta(s, a_D) - y\right)^2\right], \tag{3}$$

where y denotes a bootstrapped target constructed from the Bellman equation.

3 PAIRED RELATIVE-VALUE REGULARIZATION

3.1 POLICY-LOCAL PAIRED RELATIVE VALUES

For each transition sampled from the offline dataset, we consider the data action a_D and the current policy action $a_\pi = \pi_\phi(s)$ at the same state s. We construct a small set of policy-local candidates $\mathcal{A}_{\mathrm{loc}}(s) = \{a_1, \ldots, a_K\}$ around a_π, where $a_1 = a_\pi$. For the remaining candidates, we sample random unit directions u_k and set $a_k = \mathrm{clip}(a_\pi + \rho u_k)$, where ρ is the perturbation radius and clipping enforces the action bounds.

We use the data action a_D observed at the same state as a paired reference. For each candidate $a_k \in \mathcal{A}_{\mathrm{loc}}(s)$, we define its paired relative-value gap as

$$\Delta_k(s) = Q_\theta(s, a_k) - Q_\theta(s, a_D). \tag{4}$$

A positive Δ_k means that the critic values the candidate more highly than the paired data action, and a negative value means the opposite.

3.2 PAIRED RELATIVE-VALUE REGULARIZATION

For each state s, PaRE aggregates the paired relative-value gaps using a log-mean-exp operator,

$$z(s) = T \log \left(\frac{1}{K} \sum_{k=1}^{K} \exp\left(\frac{\Delta_k(s)}{T} \right) \right), \tag{5}$$

where $T > 0$ controls the emphasis on larger gaps. Smaller temperatures concentrate the aggregation on the largest gaps, while larger temperatures yield more uniform weighting across candidates.

The aggregated relative-value gap is then mapped to a conservative regularization term through a softplus function,

$$\mathcal{R}_{\mathrm{PaRE}}(s) = w \log\left(1 + \exp\left(\frac{z(s)}{w} \right) \right), \tag{6}$$

where $w > 0$ controls the transition width of the regularization response. The penalty increases monotonically with $z(s)$.

The log-mean-exp aggregation emphasizes candidates with larger relative-value gaps. The softplus mapping reduces the overall correction strength as the aggregated gap decreases.

3.3 THEORETICAL ANALYSIS

We analyze PaRE's relationship to policy-local value overestimation, its allocation of conservative correction, and the local value geometry captured by finite candidate sampling. We fix a state s and a policy π, and consider a single critic Q_θ. Candidate actions are held fixed when differentiating the regularizer. The analysis applies separately to each critic.

We first relate the paired value gap to candidate estimation error. Let Q^π denote the true action-value function of π, and define $e_\theta(s,a) = Q_\theta(s,a) - Q^\pi(s,a)$. For each candidate a_k,

$$e_\theta(s, a_k) = \Delta_k(s) + e_\theta(s, a_D) - A^\pi_D(s, a_k), \tag{7}$$

where $A^\pi_D(s, a_k) = Q^\pi(s, a_k) - Q^\pi(s, a_D)$ is the true value difference between the candidate and the paired data action. We use this identity to bound candidate overestimation under the following assumptions.

Assumption 1 (Local value regularity). The true value function $Q^\pi(s, \cdot)$ is L-Lipschitz on an action region containing a_D and all policy-local candidates. For any actions a, a' in this region,

$$|Q^\pi(s,a) - Q^\pi(s,a')| \le L\|a - a'\|_2. \tag{8}$$

Assumption 2 (Reference accuracy). The estimation error at the paired data action satisfies

$$|e_\theta(s, a_D)| \le \epsilon_D(s). \tag{9}$$

where $\epsilon_D(s) \ge 0$.

Proposition 1 (Policy-local overestimation bound). Under Assumptions 1 and 2, let $d_{\pi D}(s) = \|a_\pi - a_D\|_2$. If every candidate satisfies $\|a_k - a_\pi\|_2 \le \rho$, then

$$\max_{a_k \in \mathcal{A}_{\mathrm{loc}}(s)} e_\theta(s, a_k) \le \mathcal{R}_{\mathrm{PaRE}}(s) + T \log K + \epsilon_D(s) + L\big(d_{\pi D}(s) + \rho\big). \tag{10}$$

Combining Eq. 7 with $\max_k \Delta_k \le z + T \log K$ and $z \le \mathcal{R}_{\mathrm{PaRE}}$ gives the bound. Holding the other terms fixed, reducing the regularizer tightens the bound. The proof is provided in Appendix A.1.

We next examine how PaRE distributes conservative correction. The uniform distribution over candidates is $u = (1/K, \ldots, 1/K)$. Binary entropy is denoted by $H_{\mathrm{b}}(g) = -g \log g - (1-g)\log(1-g)$ for $g \in [0,1]$, with $0 \log 0 = 0$.

Proposition 2 (Variational characterization and gradient allocation). For $T, w > 0$, the PaRE regularizer admits the representation

$$\mathcal{R}_{\mathrm{PaRE}} = \max_{\substack{g \in [0,1] \\ p \in \mathcal{P}_K}} \left\{ g \sum_{k=1}^{K} p_k \Delta_k - gT D_{\mathrm{KL}}(p\|u) + w H_{\mathrm{b}}(g) \right\}, \tag{11}$$

where $\mathcal{P}_K = \{p \in \mathbb{R}^K_{\geq 0} \mid \sum_{k=1}^K p_k = 1\}$ is the probability simplex. The optimal weights are

$$p_k^* = \frac{\exp(\Delta_k/T)}{\sum_{\ell=1}^K \exp(\Delta_\ell/T)}, \qquad g^* = \sigma(z/w), \tag{12}$$

where $\sigma(x) = (1 + e^{-x})^{-1}$ is the logistic sigmoid. The gradients satisfy

$$\frac{\partial \mathcal{R}_{\mathrm{PaRE}}}{\partial \Delta_k} = g^* p_k^*, \qquad \sum_{k=1}^K \frac{\partial \mathcal{R}_{\mathrm{PaRE}}}{\partial \Delta_k} = g^*. \tag{13}$$

The shared data-action value cancels from the softmax in Eq. 12. Candidate values therefore determine the allocation p^*, while the aggregated gap relative to the data action determines the overall response g^*.

For fixed paired gaps, the candidate weight ratios and the temperature derivative of z satisfy

$$\frac{p_i^*}{p_j^*} = \exp\left(\frac{\Delta_i - \Delta_j}{T}\right), \qquad \frac{\partial z}{\partial T} = -D_{\mathrm{KL}}(p^* \| u) \leq 0. \tag{14}$$

Thus, decreasing T concentrates the weights on larger gaps and does not decrease the overall response g^* when w is fixed. Appendix A.2 provides the derivation and further analysis of the temperature limits and response sensitivity.

Let $\bar{\Delta} = K^{-1} \sum_k \Delta_k$ and $\psi_w(x) = w \log(1 + e^{x/w})$. The softmax weights in Proposition 2 yield the exact decomposition

$$z = \bar{\Delta} + T D_{\mathrm{KL}}(u \| p^*), \qquad \mathcal{R}_{\mathrm{PaRE}} = \psi_w\big(\bar{\Delta} + T D_{\mathrm{KL}}(u \| p^*)\big)\,. \tag{15}$$

The first term is the mean candidate–data value gap. The second is nonnegative, depends only on differences among candidate values, and vanishes when all candidate values are equal. This identity holds for any finite candidate set. Appendix A.2 provides the derivation and comparison with mean aggregation.

We next compare selective aggregation with mean aggregation in a local neighborhood of the policy action. Reindex the candidates as $a_k = a_\pi + \rho u_k$ for $k = 0, \ldots, K-1$, where $u_0 = 0$ and the remaining directions are sampled independently and uniformly from the unit sphere in $\mathbb{R}^{d_a}$. Assume that $Q_\theta(s, \cdot)$ is twice continuously differentiable near a_π and that no action clipping occurs. Define

$$\bar{u} = \frac{1}{K} \sum_{k=0}^{K-1} u_k, \qquad M = \frac{1}{K} \sum_{k=0}^{K-1} u_k u_k^\top, \qquad C = M - \bar{u}\bar{u}^\top, \tag{16}$$

Let z_ρ denote the log-mean-exp aggregate at radius ρ, and let $z_{\mathrm{mean},\rho} = K^{-1} \sum_{k=0}^{K-1} \Delta_k$ denote its mean-aggregation counterpart. Define

$$\mathcal{R}_\rho = \psi_w(z_\rho), \qquad \mathcal{R}_{\mathrm{mean},\rho} = \psi_w(z_{\mathrm{mean},\rho}). \tag{17}$$

At the policy action, write $g_a = \nabla_a Q_\theta(s, a_\pi)$, $\Delta_\pi = Q_\theta(s, a_\pi) - Q_\theta(s, a_D)$, and $g_0 = \sigma(\Delta_\pi / w)$.

Proposition 3 (Local comparison with mean aggregation). Under the preceding assumptions, fix $K \geq 2$ and $T, w > 0$. Conditional on the sampled directions, as $\rho \to 0$,

$$\mathcal{R}_\rho - \mathcal{R}_{\mathrm{mean},\rho} = \frac{\rho^2 g_0}{2T} g_a^\top C g_a + o(\rho^2). \tag{18}$$

Taking expectation over the sampled directions yields

$$\mathbb{E}[\mathcal{R}_\rho - \mathcal{R}_{\mathrm{mean},\rho}] = \frac{\rho^2 g_0 (K-1)^2}{2TK^2 d_a} \|g_a\|_2^2 + o(\rho^2). \tag{19}$$

Since C is positive semidefinite, the leading additional term is nonnegative and measures local value variation along the sampled directions. Its expectation is proportional to the squared action-gradient norm, with strength modulated by the policy–data value gap through g_0. Thus, in the small-radius regime, selective aggregation introduces an additional sensitivity penalty relative to mean aggregation. Appendix A.3 provides the proof, the expansion of the complete regularizer, and the extension to clipped actions.

Algorithm 1 Paired Relative-Value Regularization (PaRE)

Initialize: actor π_ϕ, critics $Q_{\theta_1}, Q_{\theta_2}$, target networks, and offline dataset $\mathcal{D}$
Parameters: PaRE parameters K, ρ, T, w, λ and actor update interval k_π

```
for each training step t do
    Sample a minibatch (s, a_D, r, s', d) ~ D
    Construct the Bellman target y using the target networks
    Construct policy-local candidates around a_π = π_φ(s)
    Compute R^(1)_PaRE(s) and R^(2)_PaRE(s) using Eqs. 4–6
    Update the critics by minimizing Eq. 20
    if t mod k_π = 0 then
        Update the actor by minimizing Eq. 62
        Update the target networks
    end if
end for
```

3.4 Practical Algorithm

We incorporate PaRE into a TD3+BC-based actor–critic framework (Fujimoto & Gu, 2021). For each critic, we evaluate the regularizer defined in Eq. 6 on the same policy-local candidate set and optimize

$$\mathcal{L}_Q = \sum_{j=1}^{2} \mathbb{E}_{\mathcal{D}} \left[\left(Q_{\theta_j}(s, a_D) - y \right)^2 \right] + \frac{\lambda}{2} \sum_{j=1}^{2} \mathbb{E}_{\mathcal{D}, \mathcal{A}_{\text{loc}}} \left[\mathcal{R}^{(j)}_{\text{PaRE}}(s) \right], \tag{20}$$

where y is the TD3 target with target-policy smoothing and the twin-critic minimum (Fujimoto et al., 2018), and λ controls the regularization strength. The TD losses are summed across critics, while the PaRE terms are averaged. Candidate actions are treated as constants during critic optimization, while gradients with respect to the critic parameters flow through both the candidate and data-action values.

The actor maximizes the twin-critic minimum with a behavior-cloning regularizer. We apply LayerNorm (Ba et al., 2016) to the critic networks. PaRE introduces no additional networks. Algorithm 1 summarizes the training procedure, and Appendix B provides the complete actor objective and implementation details.

4 Experiments

4.1 Experimental Setup

Benchmarks. We evaluate PaRE on 24 tasks from D4RL (Fu et al., 2020), covering MuJoCo locomotion and Adroit dexterous manipulation. MuJoCo includes HalfCheetah, Hopper, and Walker2d, each with Medium, Medium-Replay, Medium-Expert, and Expert datasets. Adroit includes Pen, Hammer, Door, and Relocate, each with Human, Cloned, and Expert datasets. We use the v2 datasets for MuJoCo and v1 datasets for Adroit.

Baselines. We compare PaRE with TD3+BC (Fujimoto & Gu, 2021), CQL (Kumar et al., 2020), IQL (Kostrikov et al., 2022), ReBRAC (Tarasov et al., 2023), SSAR (Luo et al., 2025), and ASPC (Jing et al., 2026), covering behavior-regularized policy learning, conservative value estimation, and in-sample value learning. We use baseline results reported in the literature (Jing et al., 2026; Tarasov et al., 2023; Luo et al., 2025).

Training and evaluation. We train PaRE for one million gradient updates on each task and report the mean and standard deviation of the final D4RL-normalized scores across three independent training seeds. PaRE-specific hyperparameters are shared across all tasks. Detailed network architectures, optimization settings, and dataset-specific behavior-cloning coefficients are provided in Appendix B.

4.2 BENCHMARK PERFORMANCE

MuJoCo locomotion. Table 1 reports the results on MuJoCo locomotion tasks. PaRE achieves the second-highest average score among the compared methods. On HalfCheetah, it obtains the highest mean scores on Medium, Medium-Replay, and Medium-Expert, and the second-highest score on Expert. It also performs competitively on Hopper, ranking first on Medium and second on Medium-Expert. On Walker2d, PaRE ranks second on Expert, while SSAR and ASPC obtain higher scores on Medium and Medium-Replay.

Table 1: Performance comparison on D4RL MuJoCo locomotion tasks. Results are reported as mean $\pm$ standard deviation. The best and second-best mean scores are highlighted in bold and underlined, respectively.

Dataset	TD3+BC	CQL	IQL	ReBRAC	SSAR	ASPC	PaRE
HalfCheetah-Medium	49.6 ± 0.2	46.9 ± 0.4	50.0 ± 0.2	65.6 ± 1.0	56.5 ± 3.7	58.7 ± 0.4	$\mathbf{68.0 \pm 0.2}$
HalfCheetah-Medium-Replay	45.8 ± 0.2	45.3 ± 0.3	42.1 ± 3.6	51.0 ± 0.8	49.6 ± 0.3	50.6 ± 0.5	$\mathbf{52.4 \pm 0.3}$
HalfCheetah-Medium-Expert	97.9 ± 1.6	95.0 ± 1.4	92.7 ± 2.8	101.1 ± 5.2	94.9 ± 1.2	99.9 ± 1.2	$\mathbf{105.5 \pm 1.0}$
HalfCheetah-Expert	100.4 ± 0.4	97.3 ± 1.1	95.5 ± 2.1	$\mathbf{105.9 \pm 1.7}$	95.5 ± 1.3	105.1 ± 1.2	105.8 ± 2.3
Hopper-Medium	62.0 ± 3.0	61.9 ± 6.4	65.2 ± 4.2	102.0 ± 1.0	101.6 ± 0.4	92.7 ± 7.2	$\mathbf{102.3 \pm 4.9}$
Hopper-Medium-Replay	47.4 ± 35.4	86.3 ± 7.3	89.6 ± 13.2	98.1 ± 5.3	$\mathbf{101.6 \pm 0.7}$	101.3 ± 0.6	100.9 ± 0.5
Hopper-Medium-Expert	103.3 ± 9.2	96.9 ± 15.1	85.5 ± 29.7	107.0 ± 6.4	103.8 ± 6.7	$\mathbf{111.0 \pm 2.1}$	108.1 ± 1.4
Hopper-Expert	108.2 ± 4.2	106.5 ± 9.1	108.8 ± 3.1	100.1 ± 8.3	109.8 ± 4.3	$\mathbf{112.3 \pm 0.4}$	100.4 ± 1.2
Walker2d-Medium	62.0 ± 3.0	79.5 ± 3.2	80.7 ± 3.4	82.5 ± 3.6	87.9 ± 2.4	$\mathbf{92.4 \pm 5.4}$	84.6 ± 1.8
Walker2d-Medium-Replay	76.6 ± 12.7	76.8 ± 10.0	75.4 ± 9.3	77.3 ± 7.9	93.5 ± 2.0	$\mathbf{97.6 \pm 0.5}$	84.0 ± 2.0
Walker2d-Medium-Expert	103.3 ± 9.2	109.1 ± 0.2	112.1 ± 0.5	111.6 ± 0.3	$\mathbf{112.5 \pm 1.4}$	111.1 ± 0.3	109.6 ± 0.2
Walker2d-Expert	108.2 ± 4.2	109.3 ± 0.1	96.9 ± 32.3	$\mathbf{112.3 \pm 0.2}$	109.6 ± 0.3	110.8 ± 0.1	111.8 ± 0.0
Average	80.4	84.2	82.9	92.9	93.1	**95.3**	94.5

Adroit dexterous manipulation. Table 2 reports the results on Adroit tasks. PaRE achieves the highest average score among the compared methods and obtains the highest mean scores on all three Pen datasets. It also ranks first on Door-Expert and second on Hammer-Expert and Relocate-Expert. Scores remain low on the Human and Cloned datasets of Door, Hammer, and Relocate for PaRE and most baselines.

Table 2: Performance comparison on D4RL Adroit dexterous manipulation tasks. Results are reported as mean $\pm$ standard deviation where available. The best and second-best mean scores are highlighted in bold and underlined, respectively.

Dataset	TD3+BC	CQL	IQL	ReBRAC	ASPC	PaRE
Pen-Human	53.8 ± 15.7	37.5	81.5 ± 17.5	103.5 ± 14.1	81.1 ± 8.1	$\mathbf{105.8 \pm 7.5}$
Pen-Cloned	71.7 ± 21.5	39.2	77.2 ± 17.7	91.8 ± 21.7	87.2 ± 4.2	$\mathbf{95.5 \pm 5.1}$
Pen-Expert	126.6 ± 24.8	107.0	133.6 ± 16.0	154.1 ± 5.4	141.2 ± 9.4	$\mathbf{155.9 \pm 2.6}$
Door-Human	0.0 ± 0.0	**9.9**	3.1 ± 2.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
Door-Cloned	0.0 ± 0.0	0.4	0.8 ± 1.0	$\mathbf{1.1 \pm 2.6}$	0.0 ± 0.0	0.2 ± 0.1
Door-Expert	81.6 ± 16.3	101.5	105.3 ± 2.8	104.6 ± 2.4	105.6 ± 0.4	$\mathbf{106.5 \pm 1.2}$
Hammer-Human	0.0 ± 0.0	**4.4**	2.5 ± 1.9	0.2 ± 0.2	2.2 ± 3.2	0.0 ± 0.0
Hammer-Cloned	0.1 ± 0.0	2.1	1.1 ± 0.5	6.7 ± 3.7	$\mathbf{12.0 \pm 9.1}$	5.3 ± 3.0
Hammer-Expert	132.8 ± 0.4	86.7	129.6 ± 71.5	$\mathbf{133.8 \pm 0.7}$	128.6 ± 0.4	133.7 ± 2.6
Relocate-Human	0.0 ± 0.0	**0.2**	0.1 ± 0.1	0.0 ± 0.0	0.1 ± 0.2	0.0 ± 0.0
Relocate-Cloned	0.0 ± 0.0	-0.1	0.2 ± 0.4	$\mathbf{0.9 \pm 1.6}$	0.0 ± 0.0	0.8 ± 0.0
Relocate-Expert	90.6 ± 18.2	95.0	106.5 ± 2.5	106.6 ± 3.2	$\mathbf{111.2 \pm 2.4}$	108.3 ± 0.3
Average	46.4	40.3	53.4	58.6	55.7	**59.3**

In Tables 1 and 2, baseline results for TD3+BC, IQL, ReBRAC, and ASPC are taken from ASPC (Jing et al., 2026), and CQL results are taken from ReBRAC (Tarasov et al., 2023). SSAR results in Table 1 are taken from its original paper (Luo et al., 2025). Standard deviations for CQL on Adroit are not reported in the source.

4.3 Ablation Study

We examine local candidates and the softplus mapping on HalfCheetah-medium using the original TD3+BC backbone. We compare full PaRE with two variants, one using only the policy–data value gap and the other replacing softplus with the linear penalty $\mathcal{R} = z$. Figure 1 shows single-seed learning curves. Removing local neighbors produces a similar learning trajectory, while the linear penalty exhibits larger fluctuations early in training. Full PaRE achieves slightly higher late-stage scores than both variants.

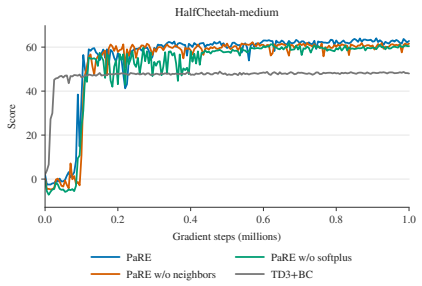


Figure 1: Single-seed ablations on HalfCheetah-medium. Removing neighbors retains only the policy action as a candidate. Removing softplus uses $\mathcal{R} = z$.

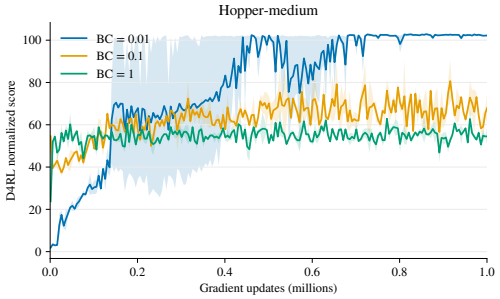


Figure 2: Sensitivity to the behavior cloning coefficient on Hopper-medium. Curves show the mean over three random seeds, with shaded regions indicating one standard deviation.

Figure 2 examines sensitivity to the behavior-cloning coefficient on Hopper-medium. The smallest tested coefficient yields the highest late-stage scores, despite slower initial learning and greater variability earlier in training. Larger coefficients learn faster initially but plateau at lower scores, highlighting the influence of behavior regularization on policy improvement.

Regularization response. Figure 4 visualizes the candidate weights p_k, state-level response g^*, and resulting correction weights $g^* p_k$ for both critics. In the displayed states, candidate weights are more concentrated on HalfCheetah-medium and Walker2d-medium than on Hopper-medium. The state-level response reduces correction for states with smaller aggregated relative-value gaps. Both critics show similar patterns, illustrating how PaRE adjusts the strength and allocation of regularization under shared hyperparameters.

4.4 Computational Cost

We compare the training time of offline RL algorithms for one million gradient updates on HalfCheetah-medium. Figure 3 reports the results.

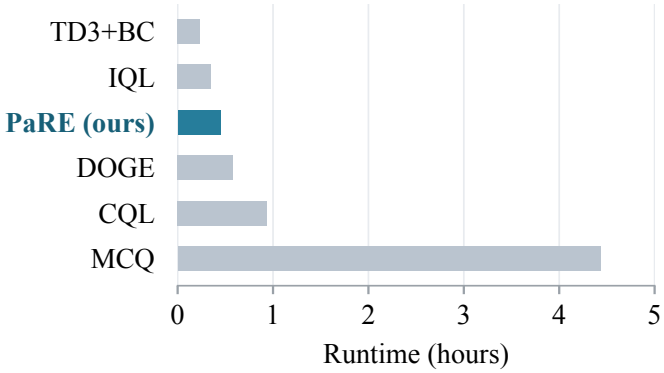
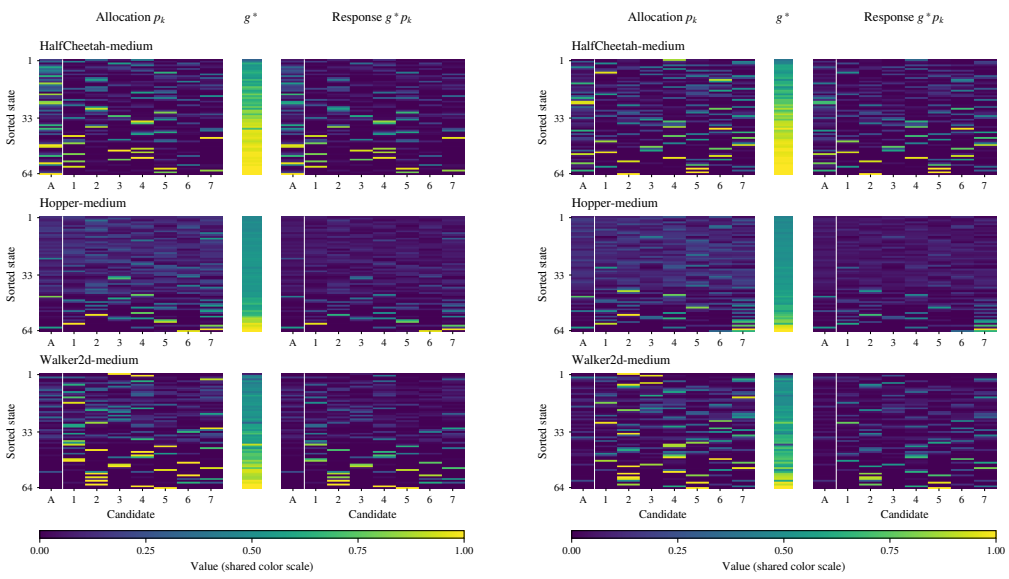


Figure 3: Runtime on HalfCheetah-medium using an NVIDIA RTX 4090 GPU. Lower is better.



(a) Critic 1 (b) Critic 2

Figure 4: Candidate allocation and regularization response for both critics on HalfCheetah-medium, Hopper-medium, and Walker2d-medium. Each row corresponds to a state. Within each task, the panels show candidate weights p_k, the state-level response g^*, and the resulting correction weights $g^* p_k$. A denotes the policy action, and 1–7 denote local candidates. All panels share the same color scale.

5 Related Work

5.1 Offline Reinforcement Learning

Policy constraints are widely used in offline RL to limit distribution shift between the learned policy and the offline data. BCQ restricts action selection through a learned generative model of behavior actions (Fujimoto et al., 2019). BEAR constrains policy updates to mitigate bootstrapping error from unsupported actions (Kumar et al., 2019), and BRAC studies behavior regularization in an actor–critic framework (Wu et al., 2019). TD3+BC introduces a simple behavior-cloning regularizer into deterministic policy optimization (Fujimoto & Gu, 2021). ReBRAC further strengthens this behavior-regularized framework through improved network architectures and training choices (Tarasov et al., 2023), and ASPC dynamically adjusts the relative scale between value maximization and behavior regularization (Jing et al., 2026). IQL avoids querying unseen actions during value learning and extracts the policy through advantage-weighted regression (Kostrikov et al., 2022). In parallel, conservative value-learning methods directly regularize value estimates for actions with limited data support (Kumar et al., 2020), which is the line of work most closely related to PaRE.

Data geometry provides another way to relax behavior constraints. DOGE learns a state-conditioned distance function to guide policy optimization toward generalizable regions (Li et al., 2023). PRDC regularizes the policy toward the action of a nearest state–action pair retrieved from the dataset (Ran et al., 2023). These methods use geometric information to guide policy updates, whereas PaRE regularizes critic values over candidates centered on the current policy action.

5.2 Conservative Value Regularization

Conservative value learning mitigates value overestimation by regularizing value estimates for actions with limited support in the offline dataset. CQL encourages conservative value estimates by penalizing action values relative to those of dataset actions (Kumar et al., 2020). MCQ assigns pseudo Q-values to out-of-distribution actions to preserve useful value generalization (Lyu et al., 2022). CPQL instead incorporates a multi-step operator into conservative value learning to exploit offline

trajectories and mitigate excessive pessimism (Kim & hwan Oh, 2026). PaRE retains a single-step TD target and adds a policy-local critic regularizer.

Other approaches control conservatism through confidence or uncertainty. CCVL conditions value functions on confidence levels to represent different degrees of conservatism (Hong et al., 2023). EDAC uses a diversified Q-ensemble to support uncertainty-based conservative value estimation (An et al., 2021). PaRE derives its correction strength and candidate weights from paired relative values within a twin-critic architecture.

Recent methods refine conservative regularization using local structure or additional information about data support. SQOG promotes smooth value generalization in neighborhoods of dataset actions (Yao et al., 2025), while SSAR adapts regularization according to state-level data quality (Luo et al., 2025). DOSER identifies out-of-distribution actions for selective regularization (Wang et al., 2026). PaRE instead constructs candidates around the current policy action and compares their values with that of the data action at the same state. The resulting paired relative values determine both the overall regularization strength and its allocation across candidates.

6 Conclusion

We proposed PaRE, which regularizes policy-local candidate values relative to the data-action value. Our analysis relates the objective to local value overestimation and shows how paired gaps determine the strength and allocation of conservative correction. Experiments on D4RL locomotion and dexterous manipulation benchmarks show competitive performance, particularly on HalfCheetah and Pen.

AI Use Statement

We used generative AI tools to improve the clarity and language of the manuscript, assist with implementing and revising experimental code, and support theoretical derivations and proof development. The authors take full responsibility for the final manuscript, theoretical claims, implementation, and reported results.

References

Gaon An, Seungyong Moon, Jang-Hyun Kim, and Hyun Oh Song. Uncertainty-based offline reinforcement learning with diversified q-ensemble. In *Advances in Neural Information Processing Systems*, pp. 7436–7447, 2021.

Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *CoRR*, abs/1607.06450, 2016.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4RL: datasets for deep data-driven reinforcement learning. *CoRR*, abs/2004.07219, 2020.

Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 20132–20145, 2021.

Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1582–1591, 2018.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2052–2062, 2019.

Joey Hong, Aviral Kumar, and Sergey Levine. Confidence-conditioned value functions for offline reinforcement learning. In *International Conference on Learning Representations*, 2023.

Tan Jing, Xiaorui Li, Chao Yao, Xiaojuan Ban, Yuetong Fang, Renjing Xu, and Zhaolin Yuan. Adaptive scaling of policy constraints for offline reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026.

Byeongchan Kim and Min hwan Oh. Peng's Q(λ) for conservative value estimation in offline reinforcement learning. In *International Conference on Learning Representations*, 2026.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. In *International Conference on Learning Representations*, 2022.

Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. In *Advances in Neural Information Processing Systems*, pp. 11761–11771, 2019.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, 2020.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *CoRR*, abs/2005.01643, 2020.

Jianxiong Li, Xianyuan Zhan, Haoran Xu, Xiangyu Zhu, Jingjing Liu, and Ya-Qin Zhang. When data geometry meets deep function: Generalizing offline reinforcement learning. In *International Conference on Learning Representations*, 2023.

Qin-Wen Luo, Ming-Kun Xie, Ye-Wen Wang, and Sheng-Jun Huang. Learning to trust bellman updates: Selective state-adaptive regularization for offline RL. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 2025.

Jiafei Lyu, Xiaoteng Ma, Xiu Li, and Zongqing Lu. Mildly conservative q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, 2022.

Yuhang Ran, Yi-Chen Li, Fuxiang Zhang, Zongzhang Zhang, and Yang Yu. Policy regularization with dataset constraint for offline reinforcement learning. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28701–28717, 2023.

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin A. Riedmiller. Deterministic policy gradient algorithms. In *International Conference on Machine Learning*, volume 32 of *JMLR Workshop and Conference Proceedings*, pp. 387–395, 2014.

Denis Tarasov, Vladislav Kurenkov, Alexander Nikulin, and Sergey Kolesnikov. Revisiting the minimalist approach to offline reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023.

Qingjun Wang, Hongtu Zhou, Hang Yu, Junqiao Zhao, Yanping Zhao, Chen Ye, Ziqiao Wang, and Guang Chen. Beyond penalization: Diffusion-based out-of-distribution detection and selective regularization in offline reinforcement learning. *CoRR*, abs/2605.08202, 2026.

Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. *CoRR*, abs/1911.11361, 2019.

Qingmao Yao, Zhichao Lei, Tianyuan Chen, Ziyue Yuan, Xuefan Chen, Jianxiang Liu, Faguo Wu, and Xiao Zhang. Offline RL with smooth OOD generalization in convex hull and its neighborhood. In *International Conference on Learning Representations*, 2025.

A Additional Theoretical Analysis

A.1 Proof of Proposition 1

We prove the policy-local overestimation bound stated in Proposition 1. The paired relative-value gap for a candidate $a_k \in \mathcal{A}_{\mathrm{loc}}(s)$ is defined in Eq. 4.

Clipping is the Euclidean projection onto the action box containing a_π, so nonexpansiveness gives $\|a_k - a_\pi\|_2 \le \rho$. Let $d_{\pi D}(s) = \|a_\pi - a_D\|_2$. For any candidate, the error decomposition, the two assumptions, and the triangle inequality give

$$\begin{aligned} Q_\theta(s,a_k) - Q^\pi(s,a_k) &= \Delta_k(s) + [Q_\theta(s,a_D) - Q^\pi(s,a_D)] \\ &\quad + [Q^\pi(s,a_D) - Q^\pi(s,a_k)] \\ &\le \Delta_k(s) + \epsilon_D(s) + L\|a_k - a_D\|_2 \\ &\le \Delta_k(s) + \epsilon_D(s) + L\big(\|a_k - a_\pi\|_2 + d_{\pi D}(s)\big) \\ &\le \Delta_k(s) + \epsilon_D(s) + L\big(\rho + d_{\pi D}(s)\big). \end{aligned} \tag{21}$$

Let $m_{\max}(s) = \max_k \Delta_k(s)$. The aggregation contains a term equal to $\exp(m_{\max}(s)/T)$, and softplus satisfies $w\log(1+e^{z/w}) \ge z$. Hence

$$\begin{aligned} z(s) &\ge T\log\left(\frac{1}{K}e^{m_{\max}(s)/T}\right) = m_{\max}(s) - T\log K, \\ \max_k \Delta_k(s) &\le z(s) + T\log K \le \mathcal{R}_{\mathrm{PaRE}}(s) + T\log K. \end{aligned} \tag{22}$$

Maximizing Eq. 21 over the candidates and applying Eq. 22 proves Eq. 10.

A.2 Proof of Proposition 2

We prove the variational representation and the gradient decomposition of PaRE stated in Proposition 2. For clarity, we omit the state argument s when no ambiguity arises and write $\Delta_k = \Delta_k(s)$ and $z = z(s)$.

Variational form of the log-mean-exp aggregation. Let $u = (1/K, \dots, 1/K)$ and let $\mathcal{P}_K$ denote the probability simplex. Expanding the KL term gives

$$\begin{aligned} F(p) &= \sum_{k=1}^K p_k\Delta_k - TD_{\mathrm{KL}}(p\|u) \\ &= \sum_{k=1}^K p_k\Delta_k - T\sum_{k=1}^K p_k\log p_k - T\log K. \end{aligned} \tag{23}$$

Introduce the Lagrangian $\mathcal{L}(p,\nu) = F(p) + \nu(\sum_k p_k - 1)$. The interior stationarity condition and normalization yield

$$\begin{aligned} 0 = \partial_{p_k}\mathcal{L} &= \Delta_k - T(\log p_k + 1) + \nu, \\ p_k &\propto e^{\Delta_k/T}, \\ p_k^* &= \frac{e^{\Delta_k/T}}{\sum_j e^{\Delta_j/T}}. \end{aligned} \tag{24}$$

For any $p \in \mathcal{P}_K$, substituting $\log p_k^* = \Delta_k/T - \log\sum_j e^{\Delta_j/T}$ gives

$$\begin{aligned} F(p) &= T\log\sum_k e^{\Delta_k/T} - T\log K - TD_{\mathrm{KL}}(p\|p^*) \\ &= z - TD_{\mathrm{KL}}(p\|p^*). \end{aligned} \tag{25}$$

KL divergence is nonnegative and vanishes only at $p = p^*$, establishing global optimality and uniqueness, including on the simplex boundary. Thus $\max_{p\in\mathcal{P}_K} F(p) = F(p^*) = z$.

Variational form of the softplus mapping. For fixed z, let $G(g) = gz + wH_{\mathrm{b}}(g)$ on $[0,1]$, with $0 \log 0 = 0$. Differentiation gives

$$\begin{aligned} G'(g) &= z + w \log \frac{1-g}{g}, \\ G''(g) &= -\frac{w}{g(1-g)} < 0, \qquad g \in (0,1), \\ G'(g^*) = 0 \quad &\Longleftrightarrow \quad \log \frac{g^*}{1-g^*} = \frac{z}{w} \quad \Longleftrightarrow \quad g^* = \sigma(z/w). \end{aligned} \tag{26}$$

The first derivative tends to $+\infty$ and $-\infty$ at the respective endpoints. Strict concavity makes the stationary point the unique global maximizer. Substituting $\log g^* = z/w - \log(1 + e^{z/w})$ and $\log(1-g^*) = -\log(1+e^{z/w})$ gives

$$\begin{aligned} \max_{g\in[0,1]} \{gz + wH_{\mathrm{b}}(g)\} &= g^* z - wg^* \log g^* - w(1-g^*)\log(1-g^*) \\ &= w \log(1 + e^{z/w}) \\ &= \mathcal{R}_{\mathrm{PaRE}}. \end{aligned} \tag{27}$$

Joint variational form of PaRE.

Substituting Eq. 25 into Eq. 27 gives

$$\mathcal{R}_{\mathrm{PaRE}} = \max_{g\in[0,1]} \left[g \max_{p\in\mathcal{P}_K} \left\{ \sum_{k=1}^{K} p_k \Delta_k - TD_{\mathrm{KL}}(p\|u) \right\} + wH_{\mathrm{b}}(g) \right]. \tag{28}$$

Since $g \geq 0$, multiplication by g preserves the maximization over p, which yields the joint variational form stated in Eq. 11. The corresponding optimal variables are given in Eq. 12.

Gradient decomposition. Differentiating the two mappings and applying the chain rule yields

$$\begin{aligned} \frac{\partial z}{\partial \Delta_k} &= \frac{e^{\Delta_k/T}}{\sum_j e^{\Delta_j/T}} = p_k^*, \\ \frac{\partial \mathcal{R}_{\mathrm{PaRE}}}{\partial z} &= \sigma(z/w) = g^*, \\ \frac{\partial \mathcal{R}_{\mathrm{PaRE}}}{\partial \Delta_k} &= \frac{\partial \mathcal{R}_{\mathrm{PaRE}}}{\partial z} \frac{\partial z}{\partial \Delta_k} = g^* p_k^*, \\ \sum_k \frac{\partial \mathcal{R}_{\mathrm{PaRE}}}{\partial \Delta_k} &= g^* \sum_k p_k^* = g^* \in (0,1). \end{aligned} \tag{29}$$

Thus, g^* determines the overall gap-level response and p_k^* allocates it across candidates, proving Eq. 13.

Temperature and width. Hold the paired gaps fixed and write $Z(T) = K^{-1} \sum_k \exp(\Delta_k/T)$, so $z(T) = T \log Z(T)$. Direct differentiation gives

$$\frac{\partial z}{\partial T} = \log Z - \frac{1}{T} \sum_k p_k^* \Delta_k. \tag{30}$$

Since $\log(p_k^*/u_k) = \Delta_k/T - \log Z$, the right-hand side equals $-D_{\mathrm{KL}}(p^*\|u)$, proving Eq. 14. The ratio p_i^*/p_j^* follows immediately from the softmax expression. As $T \to 0^+$, the inequalities

$$\max_k \Delta_k - T \log K \leq z(T) \leq \max_k \Delta_k \tag{31}$$

give the maximum limit. As $T \to \infty$, finite fixed gaps give

$$Z(T) = 1 + \frac{1}{KT} \sum_k \Delta_k + O(T^{-2}), \qquad T \log Z(T) = \frac{1}{K} \sum_k \Delta_k + O(T^{-1}), \tag{32}$$

proving the mean limit. Finally, differentiating $g^* = \sigma(z/w)$ at fixed w yields $\mathrm{d}g^*/\mathrm{d}z = g^*(1 - g^*)/w$.

Paired output derivatives. When the data-action value and candidate values are regarded as output coordinates,

$$\frac{\partial \mathcal{R}_{\mathrm{PaRE}}}{\partial Q_\theta(s,a_k)} = g^* p_k^*, \qquad \frac{\partial \mathcal{R}_{\mathrm{PaRE}}}{\partial Q_\theta(s,a_D)} = -g^*. \tag{33}$$

The reference therefore participates in the paired correction rather than acting as a detached target. These output-coordinate derivatives induce the parameter gradient

$$\nabla_\theta \mathcal{R}_{\mathrm{PaRE}} = g^* \sum_k p_k^* \left[\nabla_\theta Q_\theta(s,a_k) - \nabla_\theta Q_\theta(s,a_D)\right]. \tag{34}$$

Shared network parameters couple the outputs. The decomposition describes the regularization signal rather than a guarantee that every individual output decreases after an optimizer step.

Shared reference and equal temperatures. Subtracting the same data-action value from every candidate leaves p^* unchanged. It shifts z and hence changes the total response g^*. When $T = w = t > 0$, the two mappings also have the exact combined form

$$\mathcal{R}_{\mathrm{PaRE}} = t \log \left(e^{Q_\theta(s,a_D)/t} + \frac{1}{K} \sum_k e^{Q_\theta(s,a_k)/t} \right) - Q_\theta(s,a_D). \tag{35}$$

This is an equivalent representation of the same objective, not an additional algorithmic component.

Exact comparison with mean aggregation. Let $\bar{\Delta} = K^{-1} \sum_k \Delta_k$ and $h = z - \bar{\Delta}$. The softmax expression gives

$$\begin{aligned} T D_{\mathrm{KL}}(u \| p^*) &= \frac{T}{K} \sum_k \left[-\log K - \frac{\Delta_k}{T} + \log \sum_\ell e^{\Delta_\ell / T} \right] \\ &= T \log \left(\frac{1}{K} \sum_\ell e^{\Delta_\ell / T} \right) - \bar{\Delta} = z - \bar{\Delta} = h \geq 0. \end{aligned} \tag{36}$$

This proves Eq. 15. Equality holds exactly when all candidate gaps are equal. Subtracting a shared data-action value leaves p^* and h unchanged, while shifting $\bar{\Delta}$.

Define $\mathcal{R}_{\mathrm{mean}} = \psi_w(\bar{\Delta})$. Since $\psi_w'(x) = \sigma(x/w)$ is increasing,

$$\begin{aligned} \mathcal{R}_{\mathrm{PaRE}} - \mathcal{R}_{\mathrm{mean}} &= \int_{\bar{\Delta}}^{\bar{\Delta}+h} \sigma(x/w)\, \mathrm{d}x, \\ \sigma(\bar{\Delta}/w)\, h \leq \mathcal{R}_{\mathrm{PaRE}} - \mathcal{R}_{\mathrm{mean}} &\leq \sigma(z/w)\, h. \end{aligned} \tag{37}$$

These identities hold for arbitrary finite candidate values, without smoothness or interior-action assumptions.

A.3 Proof of Proposition 3

Fix s and write $f(a) = Q_\theta(s,a)$, $a_0 = a_\pi$, $g_a = \nabla_a f(a_0)$, and $H_a = \nabla_a^2 f(a_0)$. The data-action reference is held fixed. We first condition on the sampled directions, including $u_0 = 0$.

Finite-candidate expansion. Twice continuous differentiability gives

$$f(a_0 + \rho u_k) - f(a_0) = \rho b_k + \frac{\rho^2}{2} c_k + r_\rho(u_k), \tag{38}$$

where $b_k = g_a^\top u_k$ and $c_k = u_k^\top H_a u_k$. The integral Taylor remainder satisfies

$$\sup_{\|u\|_2 \leq 1} |r_\rho(u)| \leq \frac{\rho^2}{2} \sup_{\|v\|_2 \leq \rho} \|\nabla^2 f(a_0 + v) - H_a\|_{\mathrm{op}} = o(\rho^2). \tag{39}$$

Continuity of the Hessian on a sufficiently small closed ball gives uniformity over all candidate directions, including $u_0 = 0$. Let $\bar{b} = K^{-1} \sum_k b_k$, $\bar{c} = K^{-1} \sum_k c_k$, and $\overline{b^2} = K^{-1} \sum_k b_k^2$.

Factoring $\exp(\Delta_\pi/T)$ out of the aggregation, expanding the exponential, and using $\log(1+x) = x - x^2/2 + o(x^2)$ yields

$$\begin{aligned} z_\rho &= \Delta_\pi + T \log \left[\frac{1}{K} \sum_{k=0}^{K-1} \exp\left(\frac{\rho b_k}{T} + \frac{\rho^2 c_k}{2T} + o(\rho^2) \right) \right] \\ &= \Delta_\pi + T \log \left[1 + \frac{\rho \bar{b}}{T} + \frac{\rho^2 \bar{c}}{2T} + \frac{\rho^2 \overline{b^2}}{2T^2} + o(\rho^2) \right] \\ &= \Delta_\pi + \rho \bar{b} + \frac{\rho^2}{2} \bar{c} + \frac{\rho^2}{2T} \left(\overline{b^2} - \bar{b}^2 \right) + o(\rho^2). \end{aligned} \tag{40}$$

By Eq. 16, $\bar{b} = g_a^\top \bar{u}$, $\bar{c} = \operatorname{tr}(H_a M)$, and $\overline{b^2} - \bar{b}^2 = g_a^\top C g_a$. Thus

$$z_\rho = \Delta_\pi + \rho g_a^\top \bar{u} + \frac{\rho^2}{2} \operatorname{tr}(H_a M) + \frac{\rho^2}{2T} g_a^\top C g_a + o(\rho^2). \tag{41}$$

Comparison with mean aggregation. Averaging Eq. 38 directly yields

$$z_{\mathrm{mean},\rho} = \Delta_\pi + \rho g_a^\top \bar{u} + \frac{\rho^2}{2} \operatorname{tr}(H_a M) + o(\rho^2). \tag{42}$$

Subtracting this expression from Eq. 41 gives

$$z_\rho - z_{\mathrm{mean},\rho} = \frac{\rho^2}{2T} g_a^\top C g_a + o(\rho^2). \tag{43}$$

Here $C = K^{-1} \sum_k (u_k - \bar{u})(u_k - \bar{u})^\top \succeq 0$. Both aggregates converge uniformly to Δ_π. Applying the integral identity in Eq. 37 therefore yields

$$\begin{aligned} \mathcal{R}_\rho - \mathcal{R}_{\mathrm{mean},\rho} &= \int_{z_{\mathrm{mean},\rho}}^{z_\rho} \sigma(x/w)\, \mathrm{d}x \\ &= (g_0 + o(1))(z_\rho - z_{\mathrm{mean},\rho}) \\ &= \frac{\rho^2 g_0}{2T} g_a^\top C g_a + o(\rho^2). \end{aligned} \tag{44}$$

This proves Eq. 18. The difference is also exactly nonnegative at any radius by Eq. 37. Since $u_0 = 0$, all centered directions lie in the span of $u_1, \ldots, u_{K-1}$, giving $\operatorname{rank}(C) \le \min(d_a, K-1)$. A finite realization measures variation in the sampled span.

Fluctuation induced by finite sampling. For a uniform unit direction in $\mathbb{R}^{d_a}$,

$$\mathbb{E}[u_k] = 0, \qquad \mathbb{E}[u_k u_k^\top] = I/d_a, \quad k \ge 1. \tag{45}$$

Independence and $u_0 = 0$ imply

$$\mathbb{E}[\bar{u}] = 0, \qquad \operatorname{Cov}(\bar{u}) = \frac{K-1}{K^2 d_a} I. \tag{46}$$

Therefore,

$$\mathbb{E}[\rho g_a^\top \bar{u}] = 0, \qquad \operatorname{Var}(\rho g_a^\top \bar{u}) = \rho^2 g_a^\top \operatorname{Cov}(\bar{u}) g_a = \frac{\rho^2 (K-1)}{K^2 d_a} \|g_a\|_2^2, \tag{47}$$

which describes the first-order fluctuation rather than the exact variance of the full nonlinear aggregation.

To obtain the expected difference, independence makes the cross moments vanish, so

$$\begin{aligned} \mathbb{E}[M] &= \frac{1}{K} \sum_{k=1}^{K-1} \mathbb{E}[u_k u_k^\top] = \frac{K-1}{K d_a} I, \\ \mathbb{E}[\bar{u}\bar{u}^\top] &= \frac{1}{K^2} \sum_{k,\ell=1}^{K-1} \mathbb{E}[u_k u_\ell^\top] = \frac{K-1}{K^2 d_a} I, \\ \mathbb{E}[C] &= \mathbb{E}[M] - \mathbb{E}[\bar{u}\bar{u}^\top] = \frac{(K-1)^2}{K^2 d_a} I. \end{aligned} \tag{48}$$

The uniform remainder in Eq. 39 and smoothness of the aggregation and outer mapping permit taking expectations in Eq. 44. Substituting $\mathbb{E}[C]$ proves Eq. 19, completing Proposition 3. Similarly, taking expectations in Eq. 41 gives

$$\mathbb{E}[z_\rho] = \Delta_\pi + \frac{\rho^2(K-1)}{2Kd_a}\operatorname{tr}(H_a) + \frac{\rho^2(K-1)^2}{2TK^2d_a}\|g_a\|_2^2 + o(\rho^2). \tag{49}$$

Both coefficients account for the policy action being included as one of the K candidates.

Corollary (Expected complete PaRE regularizer). Fix the state, critic, policy, and data-action reference before sampling. Under the interior smoothness conditions of Proposition 3, for fixed $K \geq 2$ and $T, w > 0$,

$$\begin{aligned}\mathbb{E}[\mathcal{R}_\rho] = \psi_w(\Delta_\pi) &+ \frac{\rho^2 g_0(K-1)}{2Kd_a}\operatorname{tr}(H_a) \\ &+ \frac{\rho^2 g_0(K-1)}{2K^2d_a}\left(\frac{K-1}{T} + \frac{1-g_0}{w}\right)\|g_a\|_2^2 + o(\rho^2).\end{aligned} \tag{50}$$

Here $g_0 = \sigma(\Delta_\pi/w)$ is the response at the center gap, distinct from the sampled response $g^* = \sigma(z_\rho/w)$.

Proof. Write Eq. 41 as

$$z_\rho = \Delta_\pi + \rho X + \rho^2 Y + o(\rho^2), \qquad X = g_a^\top \bar{u}, \qquad Y = \tfrac{1}{2}\operatorname{tr}(H_a M) + \tfrac{1}{2T} g_a^\top C g_a. \tag{51}$$

For $\psi_w(x) = w\log(1+e^{x/w})$, its derivatives at Δ_π are $\psi_w'(\Delta_\pi) = g_0$ and $\psi_w''(\Delta_\pi) = g_0(1-g_0)/w$. A second-order expansion gives

$$\mathcal{R}_\rho = \psi_w(\Delta_\pi) + g_0\rho X + \rho^2\left(g_0 Y + \frac{g_0(1-g_0)}{2w}X^2\right) + o(\rho^2). \tag{52}$$

The directions lie in a compact set, and the continuous Hessian gives a uniform Taylor remainder. The smooth outer mapping preserves this uniformity, permitting expectation of the remainder. The direction moments give

$$\begin{aligned}\mathbb{E}[X] = 0, \qquad \mathbb{E}[X^2] &= \frac{K-1}{K^2d_a}\|g_a\|_2^2, \\ \mathbb{E}[Y] &= \frac{K-1}{2Kd_a}\operatorname{tr}(H_a) + \frac{(K-1)^2}{2TK^2d_a}\|g_a\|_2^2.\end{aligned} \tag{53}$$

Taking expectations in Eq. 52 and substituting these moments proves Eq. 50.

The softplus contribution from sampling variance can also be isolated as

$$\mathbb{E}[\psi_w(z_\rho)] - \psi_w(\mathbb{E}[z_\rho]) = \frac{\rho^2 g_0(1-g_0)(K-1)}{2wK^2d_a}\|g_a\|_2^2 + o(\rho^2). \tag{54}$$

Thus, applying softplus after averaging is not equivalent to averaging the regularizer. The gradient-norm coefficient is nonnegative, but the curvature contribution may be negative. The expansion does not imply monotonicity in radius.

Corresponding population aggregation. For a uniform unit direction u, define the sphere-only population aggregation

$$\bar{z}_\rho = T\log\mathbb{E}_u\left[\exp\left(\frac{f(a_0+\rho u) - Q_\theta(s,a_D)}{T}\right)\right]. \tag{55}$$

Taylor expansion, followed by Eq. 45, gives

$$\mathbb{E}_u\left[\exp\left(\frac{f(a_0+\rho u) - f(a_0)}{T}\right)\right] = 1 + \frac{\rho^2}{2Td_a}\operatorname{tr}(H_a) + \frac{\rho^2}{2T^2d_a}\|g_a\|_2^2 + o(\rho^2). \tag{56}$$

Applying $\log(1+x) = x + o(x)$ now yields

$$\bar{z}_\rho = \Delta_\pi + \frac{\rho^2}{2d_a}\operatorname{tr}(H_a) + \frac{\rho^2}{2Td_a}\|g_a\|_2^2 + o(\rho^2). \tag{57}$$

Unlike a finite realization, this population average has no first-order directional term. For fixed sufficiently small ρ, boundedness on the local sphere and the strong law of large numbers also imply $z_\rho \to \bar{z}_\rho$ almost surely as $K \to \infty$: the single policy-action term has weight $1/K$, which vanishes in this limit.

Clipped candidate displacements. To account for clipping, let $\delta_k = \mathrm{clip}(a_\pi + \rho u_k, -1, 1) - a_\pi$ and $\delta_0 = 0$. Define

$$\bar{\delta} = \frac{1}{K}\sum_k \delta_k, \qquad M_\delta = \frac{1}{K}\sum_k \delta_k \delta_k^\top, \qquad C_\delta = M_\delta - \bar{\delta}\bar{\delta}^\top. \tag{58}$$

Assume that the critic is twice continuously differentiable on an open neighborhood of a_π, including at the action-box boundary. Nonexpansiveness of projection gives $\|\delta_k\| \le \rho$. Taylor expanding the critic at the actual displacements and repeating the exponential and logarithmic expansions yields

$$z_\rho = \Delta_\pi + g_a^\top \bar{\delta} + \tfrac{1}{2}\,\mathrm{tr}(H_a M_\delta) + \tfrac{1}{2T} g_a^\top C_\delta g_a + o(\rho^2). \tag{59}$$

No derivative of the clipping map is needed. Applying the outer Taylor expansion gives

$$\begin{aligned}\mathcal{R}_\rho = \psi_w(\Delta_\pi) + g_0 g_a^\top \bar{\delta} + \tfrac{g_0}{2}\,\mathrm{tr}(H_a M_\delta) \\ + \tfrac{g_0}{2T} g_a^\top C_\delta g_a + \tfrac{g_0(1-g_0)}{2w}(g_a^\top \bar{\delta})^2 + o(\rho^2).\end{aligned} \tag{60}$$

Clipping generally destroys the zero-mean and isotropic properties of the original directions. At an upper coordinate boundary, for example, the outward component is removed while the inward component remains. Consequently, $\mathbb{E}[\bar{\delta}]$ need not vanish, and the interior sphere moments cannot be substituted into these formulas. For a fixed strictly interior center, clipping becomes inactive at sufficiently small radius, recovering the preceding results.

Scope of the local expansion. For the implemented ReLU critic with LayerNorm, these expansions apply in neighborhoods with fixed ReLU activation patterns and positive LayerNorm numerical stabilizers. They do not cover neighborhoods crossing activation boundaries. Equations 41 and 50 use interior sphere sampling, whereas Eqs. 59 and 60 account for clipping through actual displacements. All are small-radius asymptotic statements at fixed temperatures, rather than finite-radius error guarantees. Propositions 1 and 2 do not require the smoothness assumption.

B Implementation Details

We implement PaRE in JAX using Flax and Optax, adapting the ReBRAC implementation (Tarasov et al., 2023) to a TD3+BC-based actor–critic framework. Both the actor and critics use three hidden layers with 256 units. Each critic hidden layer applies a linear transformation followed by LayerNorm and ReLU. The actor does not use LayerNorm. PaRE introduces no additional networks. For each state, we use one policy action and seven fixed-radius local perturbations as policy-local candidates. PaRE is computed independently for the two critics using raw Q-values, and candidate actions are treated as stop-gradient inputs. Unless otherwise specified, the remaining training procedure follows the TD3+BC backbone. Table 3 summarizes the main implementation settings.

Actor objective. The actor uses the minimum of the two online critics,

$$Q_{\min}(s,a) = \min_{j\in\{1,2\}} Q_{\theta_j}(s,a), \tag{61}$$

and minimizes

$$\mathcal{L}_\pi = -\lambda_Q \mathbb{E}_{s\sim\mathcal{D}}\left[Q_{\min}(s,\pi_\phi(s))\right] + \beta_{\mathrm{BC}}\mathbb{E}_{(s,a_D)\sim\mathcal{D}}\left[\sum_m (\pi_{\phi,m}(s) - a_{D,m})^2\right], \tag{62}$$

where

$$\lambda_Q = \frac{1}{\mathrm{sg}\left[\mathbb{E}_{s\sim\mathcal{D}}|Q_{\min}(s,\pi_\phi(s))|\right] + 10^{-6}}. \tag{63}$$

The expectation used to compute λ_Q is estimated on the current minibatch, and sg denotes stop-gradient. The behavior-cloning loss is summed over action dimensions and averaged over the minibatch. During actor optimization, critic parameters are held fixed while gradients propagate through the action input. Dataset-specific behavior-cloning coefficients β_{BC} are summarized in Table 4.

Table 3: Implementation and hyperparameter settings of PaRE.

Hyperparameter	MuJoCo	Adroit
TD3+BC backbone		
Actor learning rate	1×10^{-3}	3×10^{-4}
Critic learning rate	1×10^{-3}	3×10^{-4}
Mini-batch size	1024	256
Discount factor γ	0.99	0.99
Target update rate τ	0.005	0.005
Policy noise	0.2	0.2
Policy noise clipping	0.5	0.5
Policy update frequency	2	2
Architecture		
Actor hidden dimension	256	256
Actor hidden layers	3	3
Critic hidden dimension	256	256
Critic hidden layers	3	3
Critic LayerNorm	True	True
PaRE		
Number of candidates K	8	8
Perturbation radius ρ	0.3	0.3
Temperature T	0.1	0.1
Softplus scale w	0.1	0.1
PaRE coefficient λ	0.003	0.003
Action range	$[-1, 1]$	$[-1, 1]$

Bellman target and updates. For a sampled transition (s, a_D, r, s', d), target-policy smoothing constructs

$$\tilde{a}' = \text{clip}\left(\pi_{\bar{\phi}}(s') + \epsilon, -1, 1\right), \qquad \epsilon = \text{clip}(\xi, -c, c), \quad \xi \sim \mathcal{N}(0, \sigma^2_{\text{targ}} I). \tag{64}$$

The target is

$$y = r + \gamma(1-d) \min_{j \in \{1,2\}} Q_{\bar{\theta}_j}(s', \tilde{a}'), \tag{65}$$

where $\bar{\phi}$ and $\bar{\theta}_j$ are target-network parameters, $\sigma_{\text{targ}} = 0.2$, and $c = 0.5$. The target is detached during critic optimization. We update the critics at every gradient step and the actor every two critic updates. On delayed actor-update steps, target parameters are updated by Polyak averaging using the updated online parameters:

$$\bar{\theta}_j \leftarrow (1-\tau)\bar{\theta}_j + \tau\theta_j^+, \qquad \bar{\phi} \leftarrow (1-\tau)\bar{\phi} + \tau\phi^+,$$

where θ_j^+ and ϕ^+ denote the parameters after their gradient updates and $\tau = 0.005$.

Table 4: Dataset-specific behavior-regularization coefficients β_{BC}.

Dataset	β_{BC}	Dataset	β_{BC}
halfcheetah-medium-v2	.001	pen-human-v1	.10
halfcheetah-medium-replay-v2	.01	pen-cloned-v1	.05
halfcheetah-medium-expert-v2	.01	pen-expert-v1	.01
halfcheetah-expert-v2	.01	hammer-human-v1	.01
hopper-medium-v2	.01	hammer-cloned-v1	.10
hopper-medium-replay-v2	.05	hammer-expert-v1	.01
hopper-medium-expert-v2	.10	door-human-v1	.10
hopper-expert-v2	.10	door-cloned-v1	.01
walker2d-medium-v2	.05	door-expert-v1	.05
walker2d-medium-replay-v2	.05	relocate-human-v1	.10
walker2d-medium-expert-v2	.01	relocate-cloned-v1	.10
walker2d-expert-v2	.01	relocate-expert-v1	.05