Robust Policy Expansion for Offline-to-Online RL under Diverse Data Corruption

Longxiang He[1] Deheng Ye[2] Junbo Tan[1,†] Xueqian Wang[1] Li Shen[3,†]

[1]Tsinghua University [2]Tencent [3]Shenzhen Campus of Sun Yat-sen University
longxhe@gmail.com shenli6@mail.sysu.edu.cn
† Corresponding authors

November 17, 2025

清華大學深圳国际研究生院
Tsinghua Shenzhen International Graduate School

Tencent 腾讯

Robust Policy Expansion for Offline-to-Online RL under Diverse Data Corruption

Motivation & Problem Statement A robust, simple, and scalable offline-to-online method that improves policy performance even when data corruption occurs in both offline and online phases.

Figure 1: (a) Problem Statement. A schematic illustration of the O2O attack, in which both the offline pre-training phase and the online fine-tuning phase are targeted. (b) The Kurtosis Values of Policies. CLEAN means IQL is trained without attacks. In contrast, RIQL and IQL are trained on the attacked datasets.

Key Insight Attacks on various components induce heavy-tailed behavior in the policy, leading to inefficient exploration.

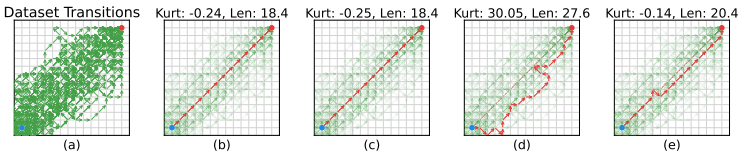

Figure 2: We study the impact of policy heavy-tailedness in the grid-world domain. An offline policy is trained using the dataset shown in Figure 2(a) and is then used to collect trajectories during the online exploration phase under both corrupted and uncorrupted settings. In Figures 2 (b)–(e), the opacity of the green arrows indicates the selection probability. Red arrows denote the most probable trajectory generated by IQL or IQL+IPW under the respective conditions. Specifically, panel (a) illustrates the dataset transitions; panels (b) and (d) show trajectories selected by IQL under clean and corrupted value functions, respectively; panels (c) and (e) show trajectories selected by IQL+IPW under clean and corrupted value functions, respectively.

Policy Expansion

To mitigate performance degradation during the transition from offline to online learning, PEX utilizes a composite policy set $\Pi = [\pi_1, \ldots, \pi_K]$ and selects actions generated by the policies in Π based on their potential utilities (e.g., critic values) in both exploration and policy learning. Specifically, for each element $\mathbb{A} = \{\boldsymbol{a}_i \sim \pi_i(\boldsymbol{s})\}$ in the policy set Π, assuming the size of Π is K, the probability of selecting $\boldsymbol{a}_i$ as the final action is

$$P_{\mathbf{w}}[i] = \frac{\exp(Q_\phi(\boldsymbol{s}, \boldsymbol{a}_i)/\alpha)}{\sum_j \exp(Q_\phi(\boldsymbol{s}, \boldsymbol{a}_j)/\alpha)}, \quad \forall i \in [1, \cdots, K], \tag{1}$$

where Q_θ denotes the offline pretrained critic function, and α represents the temperature parameter. In both PEX and our method, the policy set is defined as $\Pi = [\pi_\beta, \pi_\theta]$ with $K = 2$, where π_β is the offline pretrained policy and π_θ is the online learnable policy.

[1] Haichao Zhang and Weiwen Xu and Haonan Yu, Policy Expansion for Bridging Offline-to-Online Reinforcement Learning

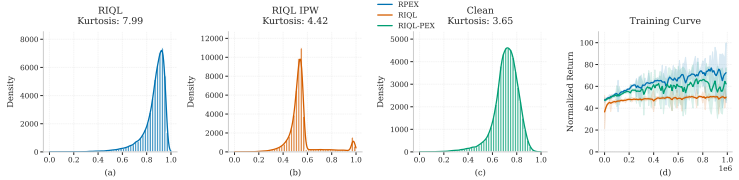

Figure 3: Action distributions generated by the offline pretrained policy under reward attack on the Halfcheetah-MR task. (a) Action distributions of RIQL under attack. (b) Action distributions of RIQL+IPW under attack. (c) Action distributions of IQL without attack. (d) Comparison of RPEX (with IPW) against RIQL-PEX (without IPW) and RIQL (Vanilla RIQL).

Method

$$P_{\mathbf{w}}[i] = \frac{\exp(Q_\phi(\boldsymbol{s}, \boldsymbol{a}_i)/\alpha) + \kappa w_{\text{ipw}}^{\pi_i}}{\sum_j \exp(Q_\phi(\boldsymbol{s}, \boldsymbol{a}_j)/\alpha) + \kappa w_{\text{ipw}}^{\pi_j}}, \quad \forall i \in [1, \cdots, K], \tag{2}$$

where

$$w_{\text{ipw}}^{\pi_i} = \text{CLIP}\left(\frac{Q_\phi - V_\psi}{\pi_i(\boldsymbol{a}_i|\boldsymbol{s})}, \text{MIN}, \text{MAX}\right). \tag{3}$$

Algorithm 1 RPEX: **R**obust **P**olicy **EX**pansion

Input: offline RL algorithm IQL or RIQL $\{L^{Q_\phi}_{\text{offline}}, L^{\pi_\beta}_{\text{offline}}\}$, online RL algorithm $\{L^{Q_\phi}_{\text{online}}, L^{\pi_\theta}_{\text{online}}\}$[2]
Initialize: UTD= M, network parameters ϕ, β, θ, corrupted offline replay buffer $\hat{\mathcal{D}}_{\text{offline}}$
Normalize the states in both the environment and the corrupted offline replay buffer $\hat{\mathcal{D}}_{\text{offline}}$
while in *offline training phase* **do**
  % offline policy training using batches from the corrupted offline replay buffer $\hat{\mathcal{D}}_{\text{offline}}$
  $\phi \leftarrow \phi - \lambda_Q \nabla_\phi L^{Q}_{\text{offline}}(\phi), \quad \beta \leftarrow \beta - \lambda_\pi \nabla_\beta L^{\pi_\beta}_{\text{offline}}(\beta)$
end while
Policy Expansion: $\tilde{\pi} = [\pi_\beta, \pi_\theta]$; transfer Q_ϕ
while in *online training phase* **do**
  for each environment step **do**
    $\boldsymbol{a}_t \sim \tilde{\pi}(\boldsymbol{a}_t|\boldsymbol{s}_t)$ according to (Eq 2),
    $\boldsymbol{s}_{t+1} \sim \mathcal{P}(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t)$,
    Attack $\{(\boldsymbol{s}_t, \boldsymbol{a}_t, r, \boldsymbol{s}_{t+1})\}$
    % Add corrupted transition into online buffer $\hat{\mathcal{D}}$
    $\hat{\mathcal{D}} \leftarrow \hat{\mathcal{D}} \cup \{(\hat{\boldsymbol{s}}_t, \hat{\boldsymbol{a}}_t, \hat{r}, \hat{\boldsymbol{s}}_{t+1})\}$
  end for
  for each gradient step **do**
    % online training using batches from both $\mathcal{D}_{\text{offline}}$ and $\mathcal{D}$
    $\phi \leftarrow \phi - \lambda_Q \nabla_\phi L^{Q}_{\text{online}}(\phi), \quad \theta \leftarrow \theta - \lambda_\pi \nabla_\theta L^{\pi_\theta}_{\text{online}}(\theta)$ for M times
    % high UTD for action corruption
  end for
end while

Main Results

Table 1: Average normalized Offline-to-Online score under random data corruption on the Medium-Replay Tasks over 5 random seeds.

Environment	Attack Element	IQL	IQL-PEX	**IQL-RPEX (ours)**	RIQL	RIQL-PEX	**RPEX (ours)**
Halfcheetah-MR	observation	21.4→21.7±5.8	21.4→21.5±2.1	21.4→**21.9**±2.8	19.73→21.3±4.2	19.73→20.9±5.3	19.73→**22.5**±3.1
	action	42.9→48.4±0.3	42.9→65.9±1.4	42.9→**69.2**±0.9	43.5→49.9±0.5	43.5→70.2±8.0	43.5→**77.8**±4.5
	reward	41.9→44.5±1.4	41.9→47.0±0.7	41.9→**52.7**±0.5	43.6→49.7±2.0	43.6→67.1 ±6.3	43.6→**73.6**±4.0
	dynamics	37.1→35.8±1.1	37.1→**37.0**±4.9	37.1→36.9±2.2	42.0→45.3±0.8	42.0→**44.7** ±0.3	42.0→44.4±0.5
Walker2d-MR	observation	8.7→17.0±6.5	8.7→20.8±4.5	8.7→**23.3**±3.3	32.2→17.0±2.2	32.2→25.6±3.2	32.2→**30.5** ±3.6
	action	64.7→**106.8**±0.9	64.7→105.3±0.6	64.7→106.4±0.5	85.9→48.8±20.9	85.9→109.2±15.6	85.9→**118.9**±10.1
	reward	77.2→90.1±9.9	77.2→90.1 ±7.2	77.2→**94.5**±6.5	81.8→91.3±1.6	81.8→91.9±1.1	81.8→**100.5**±2.9
	dynamics	14.9→4.5±1.8	14.9→**6.8**±2.5	14.9→4.7±1.2	80.0→87.4±2.7	80.0→89.5±1.3	80.0→**92.2**±1.4
Hopper-MR	observation	75.8→36.1±11.9	75.8→45.4±6.3	75.8→**76.9**±5.9	78.3→29.2±6.8	78.3→45.9±12.2	78.3→**73.1**±9.7
	action	93.4→95.0±4.4	93.4→102.2±6.4	93.4→**106.2**±5.7	75.9→95.8±3.2	75.9→93.2±10.5	75.9→**112.6**±5.4
	reward	55.2→97.8±1.2	55.2→99.8 ±2.4	55.2→**102.6**±2.8	72.9→68.3±2.9	72.9→90.1±2.9	72.9→**100.6**±2.4
	dynamics	0.8→6.0±7.5	0.8→0.7 ±0.0	0.8→**13.4**±0.9	44.6→54.2±13.6	44.6→51.0±4.5	44.6→**55.2**±4.7
Average offline score ↑		44.5	44.5	44.5	58.4	58.4	58.4
Average O2O score ↑		50.3	53.5	**59.1**	54.9	66.6	**75.16**
Average improvement percentage ↑		13.1%	20.3%	**32.7%**	-6.1%	14.1%	**28.7%**