RORL: Robust Offline Reinforcement Learning via Conservative Smoothing

Rui Yang[1]*, Chenjia Bai[2]*, Xiaoteng Ma[3], Zhaoran Wang[4], Chongjie Zhang[3], Lei Han[5]

[1]HKUST, [2]Shanghai AI Laboratory, [3]Tsinghua University, [4]Northwestern University, [5]Tencent Robotics X

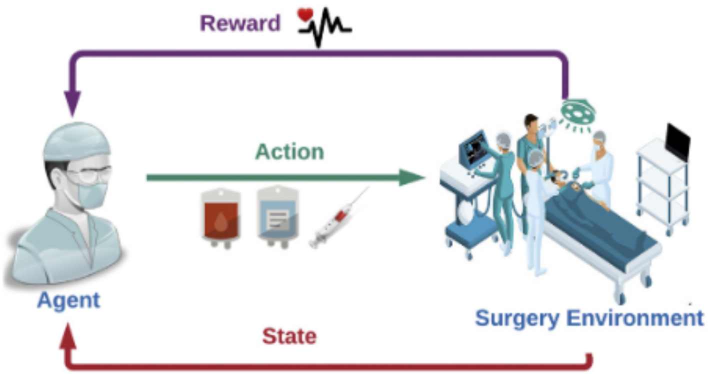
[1]HKUST, [2]Shanghai AI Laboratory, [3]Tsinghua University,

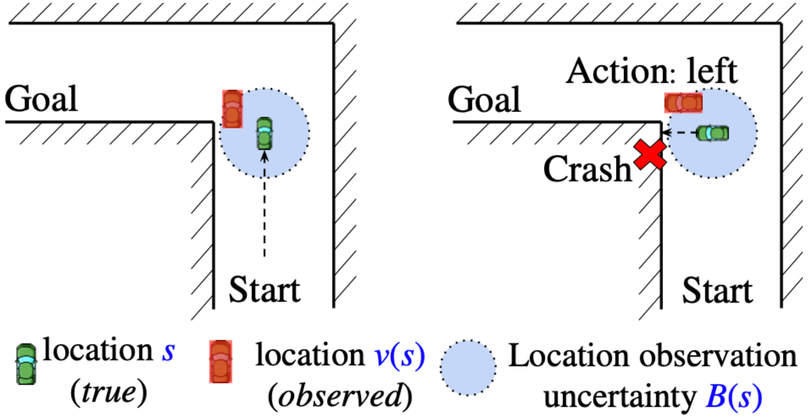
[4]Northwestern University, [5]Tencent Robotics X

Background

➤ Online interaction is costly and even prohibitive in many real-world scenarios

➤ Robustness is crucial for real-world scenarios with sensor/actuator errors and model mismatch





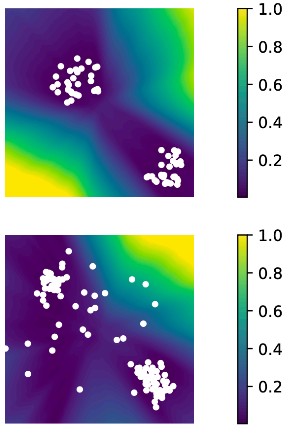
➤ **Can we learn robust policy from offline data?**

Levine S, et al. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. arXiv preprint.
Huan Zhang, et al. Robust Deep Reinforcement Learning against Adversarial Perturbations on State Observations. NeurIPS 2020.

Background

Offline RL

- PBRL: underestimating values of OOD actions according to the uncertainty estimation



(a) Uncertainty

$$\mathcal{L}_{\text{critic}} = \widehat{\mathbb{E}}_{(s,a,r,s')\sim\mathcal{D}_{\text{in}}}\left[(\widehat{\mathcal{T}}^{\text{in}}Q^k - Q^k)^2\right] + \widehat{\mathbb{E}}_{s^{\text{ood}}\sim\mathcal{D}_{\text{in}},a^{\text{ood}}\sim\pi}\left[(\widehat{\mathcal{T}}^{\text{ood}}Q^k - Q^k)^2\right],$$

$$\widehat{\mathcal{T}}^{\text{ood}}Q_\theta^k(s^{\text{ood}}, a^{\text{ood}}) := Q_\theta^k(s^{\text{ood}}, a^{\text{ood}}) - \beta_{\text{ood}}\,\mathcal{U}_\theta(s^{\text{ood}}, a^{\text{ood}}),$$

$$\mathcal{U}(s,a) := \text{Std}(Q^k(s,a)) = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(Q^k(s,a) - \bar{Q}(s,a)\right)^2}.$$

- SAC-N: increasing the number of Q networks of clipped double Q trick

$$\min_{\phi_i} \mathbb{E}_{\mathbf{s},\mathbf{a},\mathbf{s}'\sim\mathcal{D}}\left[\left(Q_{\phi_i}(\mathbf{s},\mathbf{a}) - \left(r(\mathbf{s},\mathbf{a}) + \gamma\,\mathbb{E}_{\mathbf{a}'\sim\pi_\theta(\cdot|\mathbf{s}')}\left[\min_{j=1,\ldots,N} Q_{\phi'_j}(\mathbf{s}',\mathbf{a}') - \beta\log\pi_\theta(\mathbf{a}'\mid\mathbf{s}')\right]\right)\right)^2\right]$$

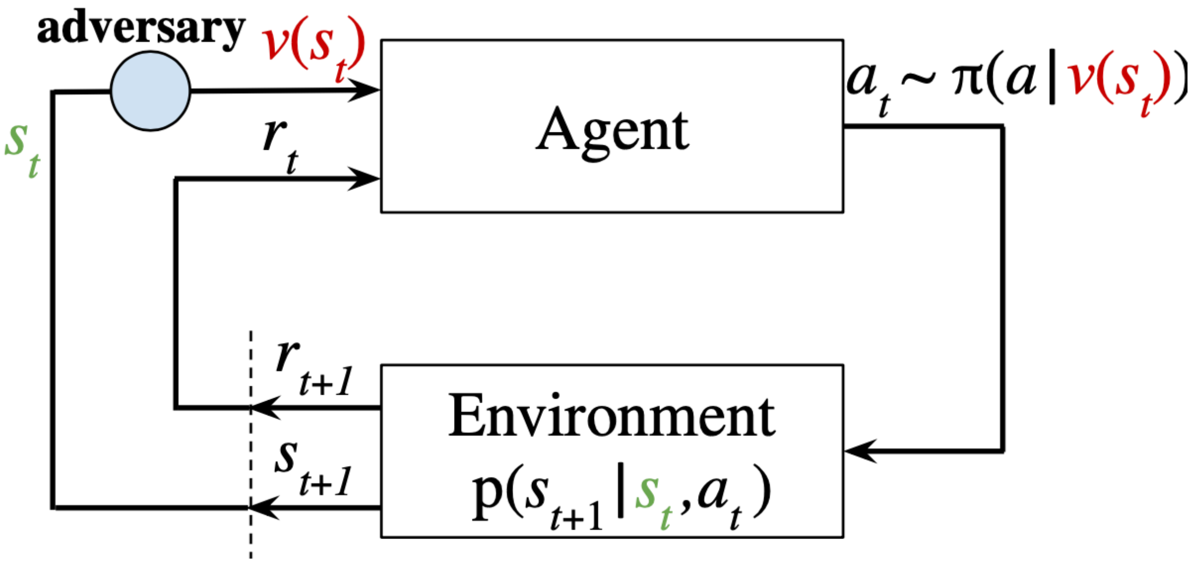
$$\max_{\theta} \mathbb{E}_{\mathbf{s}\sim\mathcal{D},\mathbf{a}\sim\pi_\theta(\cdot|\mathbf{s})}\left[\min_{j=1,\ldots,N} Q_{\phi_j}(\mathbf{s},\mathbf{a}) - \beta\log\pi_\theta(\mathbf{a}\mid\mathbf{s})\right], \tag{2}$$

Bai C, et al. Pessimistic bootstrapping for uncertainty-driven offline reinforcement learning. ICLR, 2022.
An G, et al. Uncertainty-based offline reinforcement learning with diversified q-ensemble. NeurIPS, 2021.

Background

Robust RL under adversarial attack

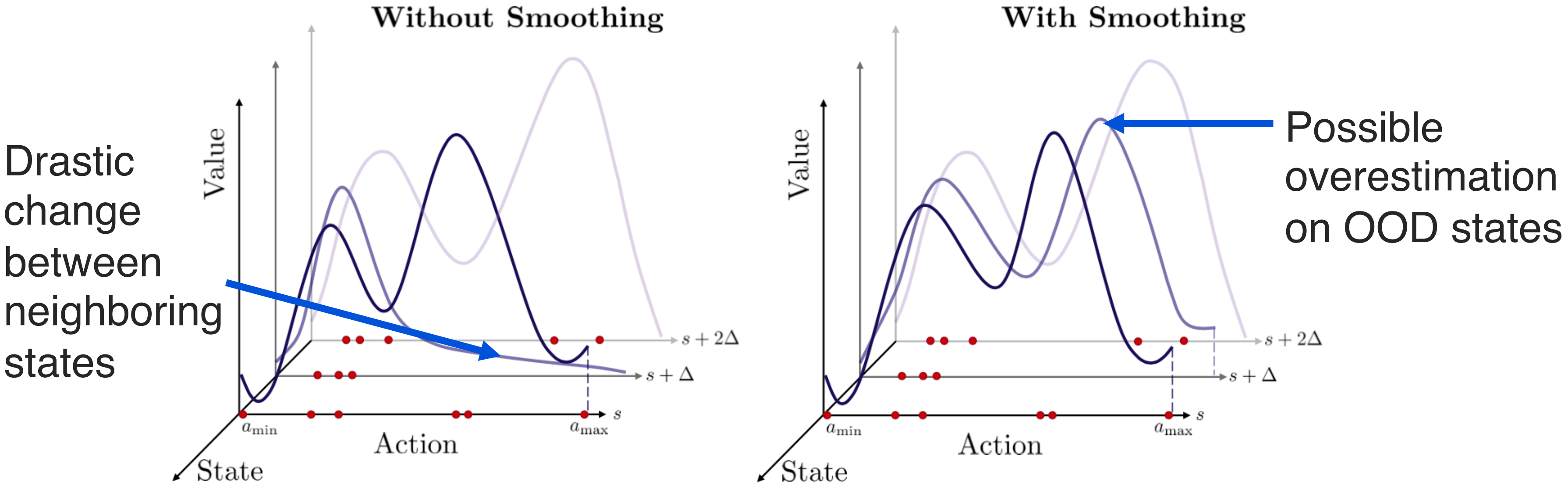
➢ Perturbation elements: observation



Huan Zhang, et al. Robust Deep Reinforcement Learning against Adversarial Perturbations on State Observations. NeurIPS 2020.

Motivating Example

Smoothing for value-based offline RL

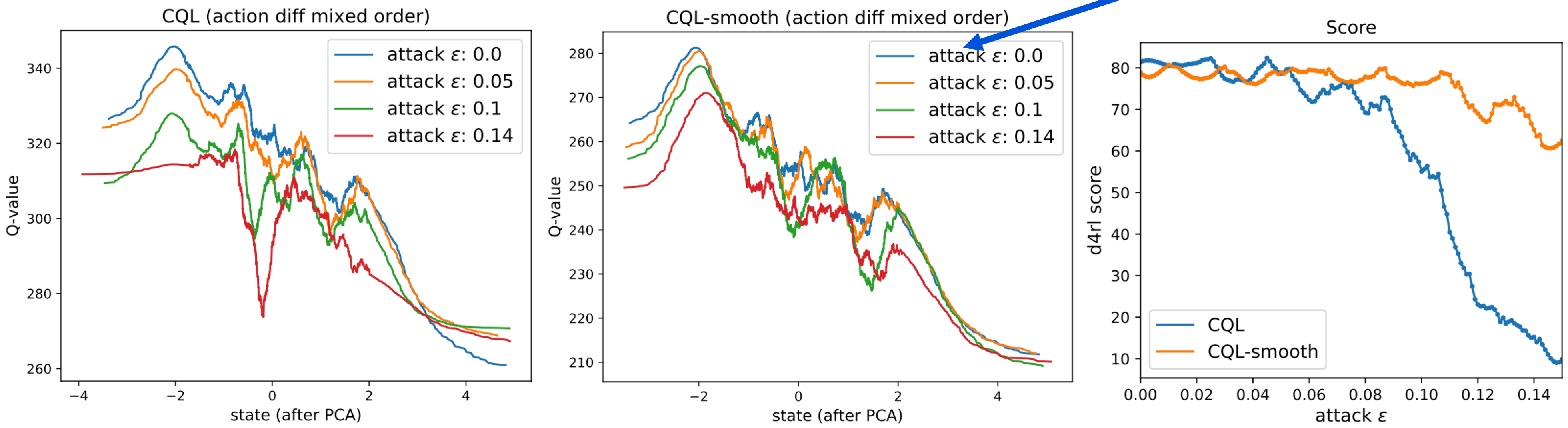


➢ <u>We need to trade off robustness and conservatism</u>

Motivating Example

Visualization for CQL

Perturbation scale, l_∞ norm



(a) Q-function of CQL (b) Q-function of CQL-smooth (c) Final performance

- CQL is susceptible to adversarial noise
- CQL-Smooth is more robust
- Robust offline RL needs to explicitly tackle potential OOD states perturbed by the attacker

Method

Robust Q function

$$\min_{\phi_i} \mathbb{E}_{s,a,r,s'\sim\mathcal{D}}\Big[\big(\widehat{\mathcal{T}}Q_{\phi_i}(s,a) - Q_{\phi_i}(s,a)\big)^2 + \beta_{\mathrm{Q}}\mathcal{L}_{\mathrm{smooth}}(s,a;\phi_i) + \beta_{\mathrm{ood}}\mathcal{L}_{\mathrm{ood}}(s;\phi_i)\Big],$$

- TD loss + smooth loss for neighbor states + underestimation for OOD states

- $\mathcal{L}_{smooth}$ is defined by:

$$\mathcal{L}_{\mathrm{smooth}}(s,a;\phi_i) = \max_{\hat{s}\in\mathbb{B}_d(s,\epsilon)} \mathcal{L}\big(Q_{\phi_i}(\hat{s},a), Q_{\phi_i}(s,a)\big)$$

$$\mathcal{L}\big(Q_{\phi_i}(\hat{s},a), Q_{\phi_i}(s,a)\big) = (1-\tau)\delta(s,\hat{s},a)_+^2 + \tau\delta(s,\hat{s},a)_-^2,$$

$$\delta(s,\hat{s},a) = Q_{\phi_i}(\hat{s},a) - Q_{\phi_i}(s,a)$$

Alleviate the overestimation of OOD states

Method

Robust Q function

$$\min_{\phi_i} \mathbb{E}_{s,a,r,s'\sim\mathcal{D}}\Big[\big(\widehat{\mathcal{T}}Q_{\phi_i}(s,a) - Q_{\phi_i}(s,a)\big)^2 + \beta_{\mathrm{Q}}\mathcal{L}_{\mathrm{smooth}}(s,a;\phi_i) + \beta_{\mathrm{ood}}\mathcal{L}_{\mathrm{ood}}(s;\phi_i)\Big],$$

- $\mathcal{L}_{ood}$ is defined by:

$$\mathcal{L}_{\mathrm{ood}}(s;\phi_i) = \mathbb{E}_{\hat{s}\sim\mathbb{B}_d(s,\epsilon),\hat{a}\sim\pi_\theta(\hat{s})}\big(\widehat{\mathcal{T}}_{\mathrm{ood}}Q_{\phi_i}(\hat{s},\hat{a}) - Q_{\phi_i}(\hat{s},\hat{a})\big)^2$$

$$\widehat{\mathcal{T}}_{\mathrm{ood}}Q_{\phi_i}(\hat{s},\hat{a}) := Q_{\phi_i}(\hat{s},\hat{a}) - \lambda u(\hat{s},\hat{a})$$

$$u(\hat{s},\hat{a}) := \sqrt{\frac{1}{K}\sum_{k=1}^{K}\big(Q_{\phi_i}(\hat{s},\hat{a}) - \bar{Q}_{\phi}(\hat{s},\hat{a})\big)^2}$$

Method

Robust Policy

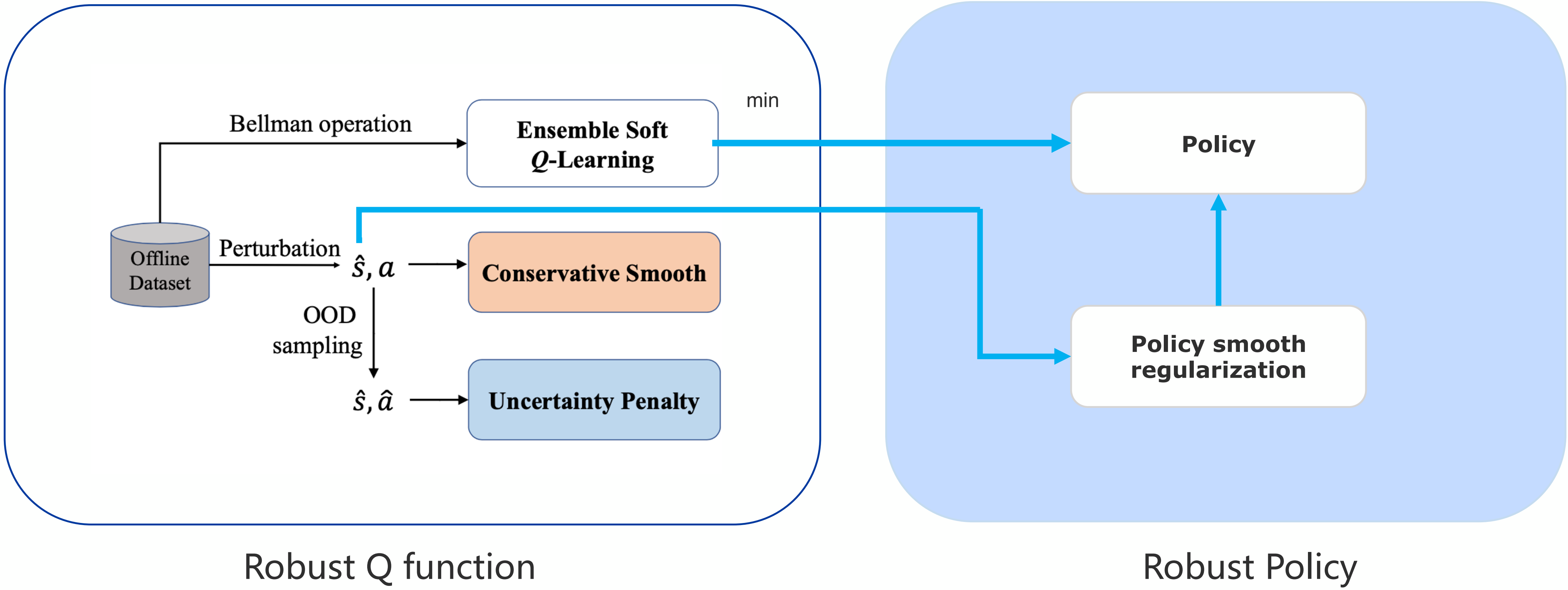
- Based on the robust and conservative value functions, we simply smooth the policy as

$$\min_{\theta} \left[\mathbb{E}_{s\sim\mathcal{D},a\sim\pi_\theta(\cdot|s)} \left[-\min_{j=1,\ldots,K} Q_{\phi_j}(s,a) + \alpha \log \pi_\theta(a|s) + \beta_{\mathrm{P}} \max_{\hat{s}\in\mathbb{B}_d(s,\epsilon)} D_{\mathrm{J}}\big(\pi_\theta(\cdot|s) \| \pi_\theta(\cdot|\hat{s})\big) \right] \right]$$

$$D_{\mathrm{J}}(P\|Q) = \tfrac{1}{2}[D_{\mathrm{KL}}(P\|Q) + D_{\mathrm{KL}}(Q\|P)]$$

Method

Overall Framework



Experiments

What are the Advantages of RORL over Previous Offline RL Algorithms?

- Performance improves on clean environments

- More robust against adversarial perturbation

Experiments

Benchmark Results

RORL only uses 10 ensemble Q networks to outperform the SOTA method EDAC with 10~50 Q networks!

Table 1: Normalized average returns on Gym tasks, averaged over 4 random seeds. Part of the results are reported in the EDAC paper. Top two scores for each task are highlighted.

Task Name	BC	CQL	PBRL	SAC-10 (Reproduced)	EDAC (Paper)	EDAC-10 (Reproduced)	RORL (Ours)
halfcheetah-random	2.2±0.0	**31.3±3.5**	11.0±5.8	**29.0±1.5**	28.4±1.0	13.4 ± 1.1	28.5±0.8
halfcheetah-medium	43.2±0.6	46.9±0.4	57.9 ±1.5	64.9±1.3	**65.9±0.6**	64.1±1.1	**66.8±0.7**
halfcheetah-medium-expert	44.0±1.6	95.0±1.4	92.3±1.1	107.1±2.0	106.3±1.9	**107.2±1.0**	**107.8±1.1**
halfcheetah-medium-replay	37.6±2.1	45.3±0.3	45.1±8.0	**63.2±0.6**	61.3±1.9	60.1±0.3	**61.9±1.5**
halfcheetah-expert	91.8±1.5	97.3±1.1	92.4±1.7	104.9±0.9	**106.8±3.4**	104.0±0.8	**105.2±0.7**
hopper-random	3.7±0.6	5.3±0.6	**26.8±9.3**	25.9±9.6	25.3±10.4	16.9±10.1	**31.4±0.1**
hopper-medium	54.1±3.8	61.9±6.4	75.3±31.2	0.8±0.2	101.6±0.6	**103.6±0.2**	**104.8±0.1**
hopper-medium-expert	53.9±4.7	96.9±15.1	**110.8±0.8**	6.1±7.7	110.7±0.1	58.1±22.3	**112.7±0.2**
hopper-medium-replay	16.6±4.8	86.3±7.3	100.6±1.0	**102.9±0.9**	101.0±0.5	**102.8±0.3**	**102.8±0.5**
hopper-expert	107.7±9.7	106.5±9.1	**110.5±0.4**	1.1±0.5	110.1±0.1	77.0±43.9	**112.8±0.2**
walker2d-random	1.3±0.1	5.4±1.7	8.1±4.4	1.5±1.1	**16.6±7.0**	6.7±8.8	**21.4±0.2**
walker2d-medium	70.9±11.0	79.5±3.2	89.6±0.7	46.7±45.3	**92.5±0.8**	87.6±11.0	**102.4±1.4**
walker2d-medium-expert	90.1±13.2	109.1±0.2	110.1±0.3	**116.7±1.9**	114.7±0.9	115.4±0.5	**121.2±1.5**
walker2d-medium-replay	20.3±9.8	76.8±10.0	77.7±14.5	89.6±3.1	87.1±2.3	**94.0±1.2**	**90.4 ± 0.5**
walker2d-expert	108.7±0.2	109.3±0.1	108.3±0.3	1.2±0.7	**115.1±1.9**	57.8±55.7	**115.4 ± 0.5**
Average	49.7	70.2	74.4	50.8	**82.9**	71.2	**85.7**
Total	746.1	1052.8	1116.5	761.6	**1243.4**	1068.7	**1285.7**

Experiments

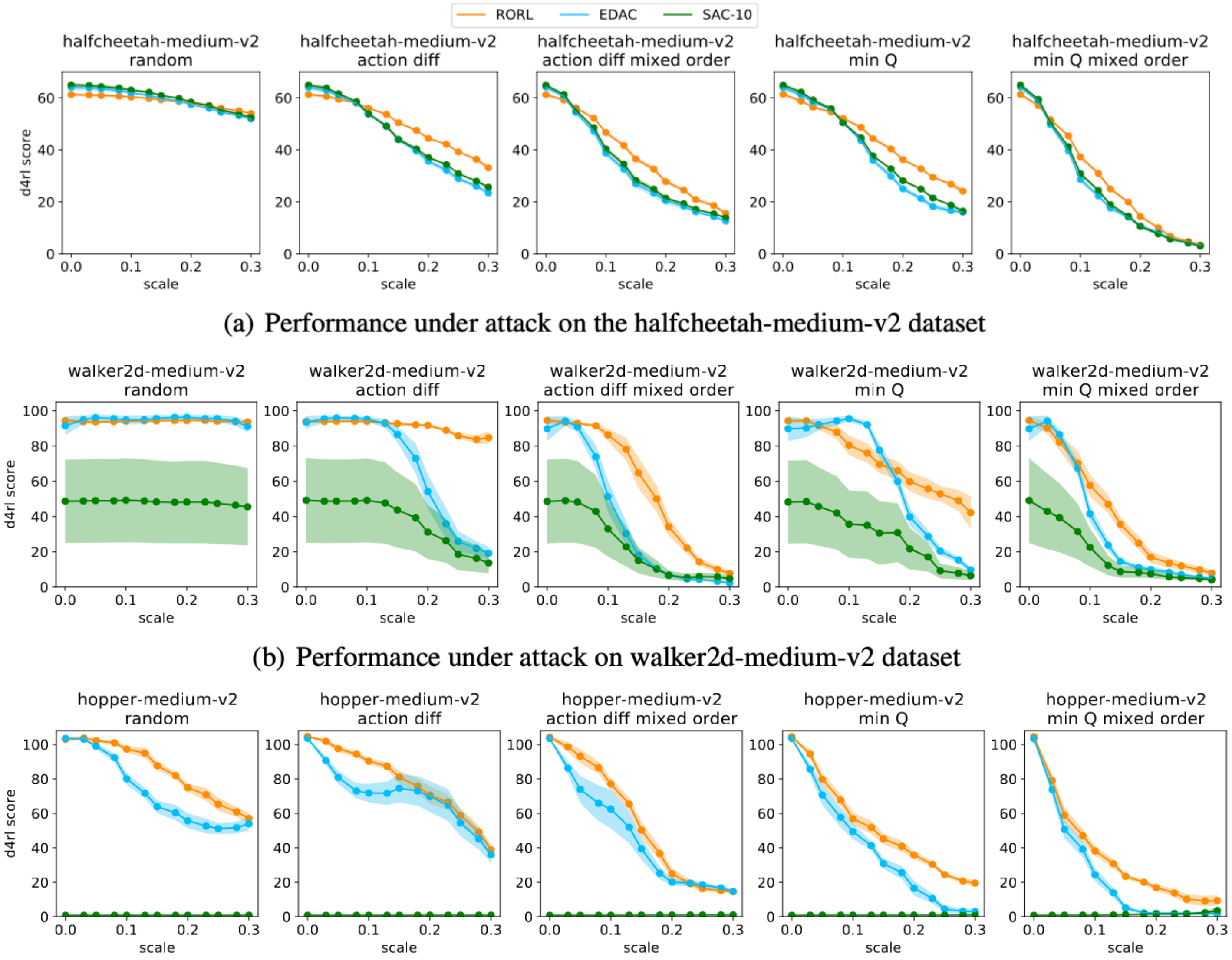
Adversarial Attack

- Attack Methods
  - Random: uniformly sampling perturbed states in an l_∞ ball of norm ϵ
  - Action diff: $\min_{\hat{s}\in\mathbb{B}_d(s,\epsilon)} -D_{\mathrm{J}}\left(\pi_\theta(\cdot|s)\|\pi_\theta(\cdot|\hat{s})\right)$
  - Min Q: $\min_{\hat{s}\in\mathbb{B}_d(s,\epsilon)} Q(s,\pi_\theta(\hat{s}))$

- Optimization
  - Zero-order: sampling 50 states and finding the minimum
  - Mixed-order: sampling 20 initial states and performing gradient decent for 10 steps with a step size of $\frac{1}{10}\epsilon$ for each initial state, and selecting the minimum

Experiments

Adversarial Attack: robustness under adversarial attack



(a) Performance under attack on the halfcheetah-medium-v2 dataset

(b) Performance under attack on walker2d-medium-v2 dataset

(c) Performance under attack on hopper-medium-v2 dataset

Experiments

Adversarial Attack: ablations of attack experiments

- ➢Each component contributes to the performance under different types of attack
- ➢The OOD loss and policy smoothing loss are more effective against attacks

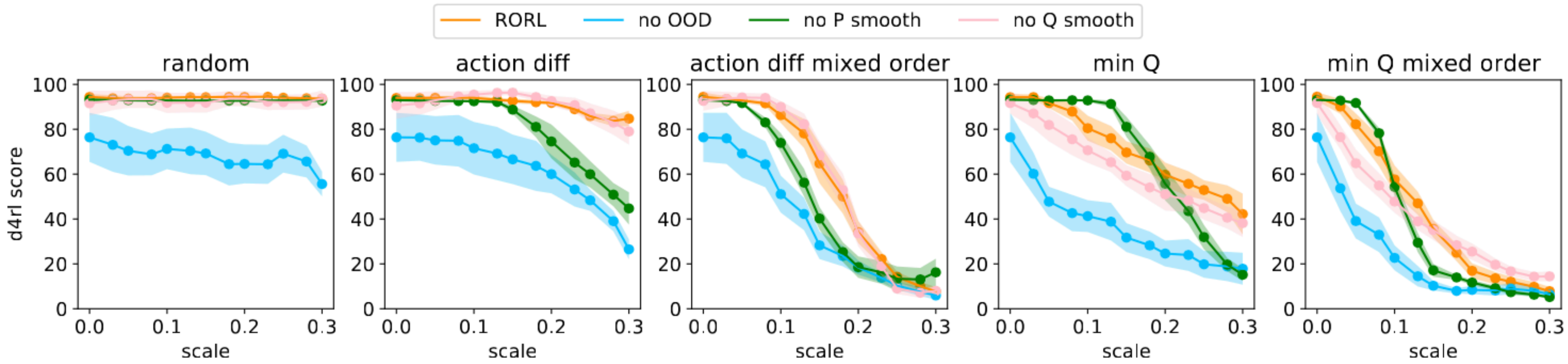


Figure 11: Ablations of different components against in the adversarial experiments.

Theoretical Analysis

RORL enjoys better property in Linear MDPs than PBRL

$$\widetilde{w}_t = \min_{w \in \mathcal{R}^d} \Big[\sum_{i=1}^{m} \big(y_t^i - Q_w(s_t^i, a_t^i)\big)^2 + \sum_{i=1}^{m} \frac{1}{|\mathbb{B}_d(s_t^i, \epsilon)|} \sum_{\hat{s}_t^i \in \mathcal{D}_{\text{ood}}(s_t^i)} \big(Q_w(s_t^i, a_t^i) - Q_w(\hat{s}_t^i, a_t^i)\big)^2 + \sum_{(\hat{s}, \hat{a}, \hat{y}) \sim \mathcal{D}_{\text{ood}}} \big(\hat{y} - Q_w(\hat{s}, \hat{a})\big)^2 \Big],$$

Theorem 2. *For all the OOD datapoint $(\hat{s}, \hat{a}, \hat{y}) \in \mathcal{D}_{\text{ood}}$, if we set $\hat{y} = \mathcal{T}V_{t+1}(s^{\text{ood}}, a^{\text{ood}})$, it then holds for $\beta_t = \mathcal{O}\big(T \cdot \sqrt{d} \cdot \log(T/\xi)\big)$ that*

$$\Gamma_t^{\text{lcb}}(s_t, a_t) = \beta_t \big[\phi(s_t, a_t)^\top \widetilde{\Lambda}_t^{-1} \phi(s_t, a_t)\big]^{1/2} \tag{18}$$

forms a valid ξ-uncertainty quantifier, where $\widetilde{\Lambda}_t$ is the covariance matrix of RORL.

Corollary (Corollary 2 restate). *Under the same conditions as Theorem 2, it holds that* $\text{SubOpt}(\pi^*, \hat{\pi}) \leq \sum_{t=1}^{T} \mathbb{E}_{\pi^*}\big[\Gamma_t^{\text{lcb}}(s_t, a_t)\big] < \sum_{t=1}^{T} \mathbb{E}_{\pi^*}\big[\Gamma_t^{\text{lcb\_PBRL}}(s_t, a_t)\big]$.

Conclusion

- We propose RORL to learn robust RL policies from offline datasets
- Specifically, we smooth the policy and the value functions of the perturbed states while adaptively underestimating their values based on uncertainty
- RORL outperforms current SOTA algorithm with fewer ensemble Q networks and is considerably robust to different types of adversarial perturbations