模式分析与机器智能
工业和信息化部重点实验室
MIIT Key Laboratory of
Pattern Analysis & Machine Intelligence

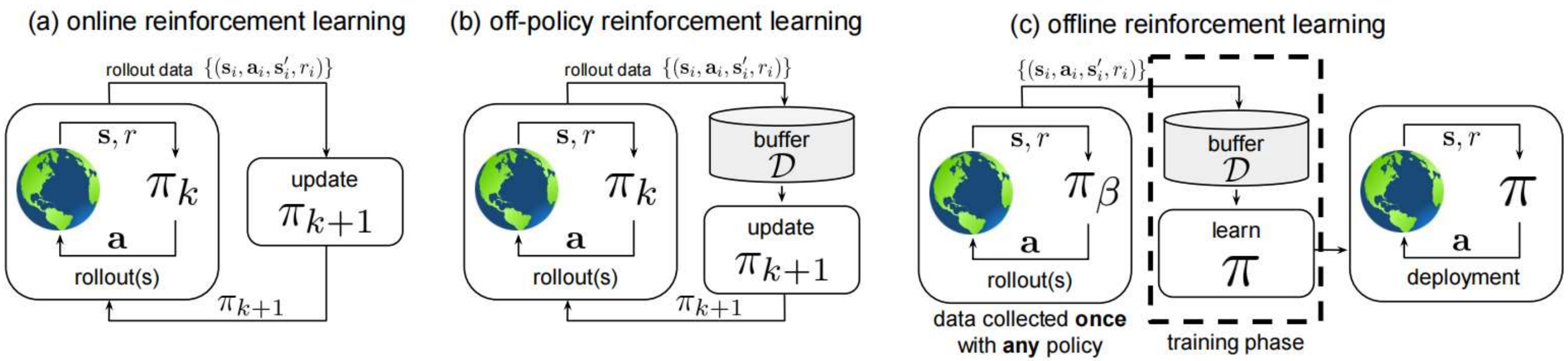
ParNeC 模式识别与神经计算研究组
PAttern Recognition and NEural Computing

Offline RL with No OOD Actions: In-Sample Learning via Implicit Value Regularization

Haoran Xu[1*] **Li Jiang**[2] **Jianxiong Li**[1] **Zhuoran Yang**[3] **Zhaoran Wang**[4]
Victor Wai Kin Chan[2] **Xianyuan Zhan**[1,5]
[1]Institute for AI Industry Research (AIR), Tsinghua University
[2]Tsinghua-Berkeley Shenzhen Institute (TBSI), Tsinghua University
[3]Yale University [4]Northwestern University [5]Shanghai Artificial Intelligence Laboratory
{ryanxhr,jiangli3859,zhanxianyuan}@gmail.com

ICLR2023

Background



MDP=$(S, A, R, \gamma, P, \rho_0)$

Objective: $J(\pi) = \mathbb{E}_{s_0 \sim \rho_0, a \sim \pi(\cdot|s), s' \sim P(\cdot|s,a)}[\sum_{t=0}^{T} \gamma^t r(s_t, a_t)]$

Policy: $\max J(\pi) \rightarrow \pi(a|s)$ $\longrightarrow$ actor

State value function: $V(s) = \mathbb{E}_{s_t, a_t \sim \rho_\pi}[\sum_{t=0}^{T} \gamma^t r(s_t, a_t) \,|s_0 = s]$

State-action value function: $Q(s, a) = \mathbb{E}_{s_t, a_t \sim \rho_\pi}[\sum_{t=0}^{T} \gamma^t r(s_t, a_t) \,|s_0 = s, a_0 = a]$

critic

policy iteration

- policy evaluation: use trajectories of π to evaluate Q and V

$$B^\pi Q(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a), a' \sim \pi(\cdot|s')}[Q(s', a')]$$

- policy improvement: use Q and V to update π

Background

Off policy evaluation(OPE)

Bellman operator $\mathcal{T}^{\pi}Q(s,a) \approx \mathbb{E}_{s'\sim\mathcal{B}}[r+\gamma Q(s',\pi(s'))]$

Extrapolation Error

- Absent Data: the estimate of $Q_\theta(s,\pi(s))$ may be arbitrarily bad without sufficient data near $(s,\pi(s))$
- Model Bias: for a stochastic MDP, without infinite state-action visitation, sampling produces a biased estimate of the transition dynamics
- Training Mismatch: due to distribution shift, there is a mismatch in training

$$\approx \frac{1}{|\mathcal{B}|}\sum_{(s,a,r,s')\in\mathcal{B}} ||r+\gamma Q_{\theta'}(s',\pi(s'))-Q_\theta(s,a)||^2$$

Because of offline setting, overestimate due to extrapolation error cannot be corrected

Background

The main idea of offline RL: Constraint the learned policy is close to the behavior policy

- Policy constraint IQL，TD3+BC

$$\pi_{k+1} = \arg\max_{\pi \in \Pi} \mathbb{E}_{\mathbf{a} \sim \pi(\cdot|\mathbf{s})}[A^{\pi_k}(\mathbf{s}, \mathbf{a})]$$
$$\text{s.t. } D_{\mathrm{KL}}(\pi(\cdot|\mathbf{s}) || \pi_\beta(\cdot|\mathbf{s})) \leq \epsilon$$

- Value Regularization CQL

$$\arg\min_Q \ \alpha \cdot \left(\mathbb{E}_{\mathbf{s} \sim \mathcal{D}, \mathbf{a} \sim \mu(\mathbf{a}|\mathbf{s})}[Q(\mathbf{s}, \mathbf{a})] - \mathbb{E}_{\mathbf{s} \sim \mathcal{D}, \mathbf{a} \sim \hat{\pi}_\beta(\mathbf{a}|\mathbf{s})}[Q(\mathbf{s}, \mathbf{a})]\right) + \frac{1}{2} \mathbb{E}_{\mathbf{s}, \mathbf{a}, \mathbf{s}' \sim \mathcal{D}}\left[\left(Q(\mathbf{s}, \mathbf{a}) - \hat{\mathcal{B}}^\pi \hat{Q}^k(\mathbf{s}, \mathbf{a})\right)^2\right]$$

- Others
  - Supervised Learning Decision transformer、RvS
  - Model based Combo
  - ......

Model free RL in offline setting: Pessimistic estimate

Background

ParNeC 模式识别与神经计算研究组 PAttern Recognition and NEural Computing

In-sample Learning

only use in-sample actions of offline dataset to evaluate policy

IQL

$$\min_V \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\left|\tau - \mathbb{1}\left(Q(s,a) - V(s) < 0\right)\right|\left(Q(s,a) - V(s)\right)^2\right]$$

$$\min_Q \mathbb{E}_{(s,a,s')\sim\mathcal{D}}\left[\left(r(s,a) + \gamma V(s') - Q(s,a)\right)^2\right],$$

Extreme Q-learning

$$\mathcal{J}(V) = \mathbb{E}_{\mathbf{s},\mathbf{a}\sim\mu}\left[e^{(\hat{Q}^k(\mathbf{s},\mathbf{a}) - V(\mathbf{s}))/\beta}\right] - \mathbb{E}_{\mathbf{s},\mathbf{a}\sim\mu}[(\hat{Q}^k(\mathbf{s},\mathbf{a}) - V(\mathbf{s}))/\beta] - 1$$

$$\mathcal{L}(\phi) = \mathbb{E}_{(\mathbf{s},\mathbf{a},\mathbf{s}')\sim\mathcal{D}}\left[(Q_\phi(\mathbf{s},\mathbf{a}) - r(\mathbf{s},\mathbf{a}) - \gamma V_\theta(\mathbf{s}'))^2\right]$$

Some other methods need to estimate behavior policy $\mu(a|s)$

In-sample learning avoids extrapolation error to get accurate Q and V on state-action pairs in offline dataset, meanwhile, in-sample learning can decouple policy improvement and policy evaluation.

Method

Behavior-regularized MDP problem

$$\max_{\pi} \mathbb{E}\Big[\sum_{t=0}^{\infty}\gamma^t\Big(r(s_t,a_t)-\alpha\cdot f\Big(\frac{\pi(a_t|s_t)}{\mu(a_t|s_t)}\Big)\Big)\Big]$$

policy evaluation operator $(\mathcal{T}_f^{\pi})Q(s,a) := r(s,a)+\gamma\mathbb{E}_{s'|s,a}\left[V(s')\right]$

$$V(s)=\mathbb{E}_{a\sim\pi}\Big[Q(s,a)-\alpha f\Big(\frac{\pi(a|s)}{\mu(a|s)}\Big)\Big].$$

Assumption 1. *Assume $\pi(a|s)>0 \Rightarrow \mu(a|s)>0$ so that π/μ is well-defined.*

Assumption 2. *Assume the function $f(x)$ satisfies the following conditions on $(0,\infty)$: (1) $f(1)=0$; (2) $h_f(x)=xf(x)$ is strictly convex; (3) $f(x)$ is differentiable.*

$h_f'(x)=f(x)+xf'(x), g_f(x)=(h_f')^{-1}(x)$

Method

Derivation

$$J(\pi)=\max_{\pi}\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^{t}r(s_t,a_t)-\alpha\cdot f\left(\frac{\pi(a_t|s_t)}{\mu(a_t|s_t)}\right)\right]=\max\ V(s)=\max\ \mathbb{E}_{a\sim\pi}\left[Q(s,a)-\alpha f\left(\frac{\pi(a|s)}{\mu(a|s)}\right)\right]$$

$$\pi(a|s)\geqslant 0,\ \sum_{a}\pi(a|s)=1$$

The Lagrangian function:

$$L(\pi,\beta,u)=\sum_{s}d^{\pi}(s)\sum_{a}\pi(a|s)\left(Q(s,a)-\alpha f\left(\frac{\pi(a|s)}{\mu(a|s)}\right)\right)$$
$$-\sum_{s}d^{\pi}(s)\left[u(s)\left(\sum_{a}\pi(a|s)-1\right)-\sum_{a}\beta(a|s)\pi(a|s)\right]$$

KKT conditions:

$$0\leq\pi(a|s)\leq 1 \text{ and } \sum_{a}\pi(a|s)=1$$
$$0\leq\beta(a|s)$$
$$\beta(a|s)\pi(a|s)=0$$
$$Q(s,a)-\alpha h_f'\left(\frac{\pi(a|s)}{\mu(a|s)}\right)-u(s)+\beta(a|s)=0$$

Method

$$Q(s,a) - \alpha h_f'\left(\frac{\pi(a|s)}{\mu(a|s)}\right) - u(s) + \beta(a|s) = 0 \longrightarrow \pi(a|s) = \mu(a|s)\cdot g_f\left(\frac{1}{\alpha}\left(Q(s,a) - u(s) + \beta(a|s)\right)\right)$$

complementary slackness $\beta(a|s)\pi(a|s) = 0 \longrightarrow \pi(a|s) = \mu(a|s)\cdot \max\left\{g_f\left(\frac{1}{\alpha}\left(Q(s,a)-u(s)\right)\right), 0\right\}$

normalization condition $\longrightarrow$ $\mathbb{E}_{a\sim\mu}\left[\max\left\{g_f\left(\frac{1}{\alpha}\left(Q(s,a)-u(s)\right)\right), 0\right\}\right] = 1$

optimal value function $\longrightarrow$

$$\begin{aligned}
V^*(s) &= \mathcal{T}_f^* V^*(s) \\
&= \sum_a \pi^*(a|s)\left(Q^*(s,a) - \alpha f\left(\frac{\pi^*(a|s)}{\mu(a|s)}\right)\right) \\
&= \sum_a \pi^*(a|s)\left(u^*(s) + \alpha \frac{\pi^*(a|s)}{\mu(a|s)} f'\left(\frac{\pi^*(a|s)}{\mu(a|s)}\right)\right) \\
&= u^*(s) + \alpha \sum_a \frac{\pi^*(a|s)^2}{\mu(a|s)} f'\left(\frac{\pi^*(a|s)}{\mu(a|s)}\right) \\
&= u^*(s) + \alpha \mathbb{E}_{a\sim\mu}\left[\left(\frac{\pi^*(a|s)}{\mu(a|s)}\right)^2 f'\left(\frac{\pi^*(a|s)}{\mu(a|s)}\right)\right]
\end{aligned}$$

Method

optimality conditions of the behavior regularized MDP

$$Q^*(s,a) = r(s,a) + \gamma \mathbb{E}_{s'|s,a}\left[V^*(s')\right]$$

$$\pi^*(a|s) = \mu(a|s) \cdot \max\left\{g_f\left(\frac{Q^*(s,a) - u^*(s)}{\alpha}\right), 0\right\}$$

$$V^*(s) = u^*(s) + \alpha \mathbb{E}_{a\sim\mu}\left[\left(\frac{\pi^*(a|s)}{\mu(a|s)}\right)^2 f'\left(\frac{\pi^*(a|s)}{\mu(a|s)}\right)\right]$$

$$\mathbb{E}_{a\sim\mu}\left[\max\left\{g_f\left(\frac{1}{\alpha}(Q^*(s,a) - u^*(s))\right), 0\right\}\right] = 1$$

zero-forcing support constraint $\mu(a|s) = 0 \Rightarrow \pi(a|s) = 0$

α-divergence: $D_\alpha(\mu,\pi) = \frac{1}{\alpha(\alpha-1)} \mathbb{E}_\pi\left[\left(\frac{\pi}{\mu}\right)^{-\alpha} - 1\right]$ $\begin{cases} \alpha = -1 & \chi^2 - \text{divergence} \\ \alpha \to 0 & \text{Reverse KL divergence} \end{cases}$

$\alpha \leqslant 0$ mode-seeking

Method

ParNeC 模式识别与神经计算研究组 PAttern Recognition and NEural Computing

Sparse Q-learning(SQL) $\alpha = -1,\ f(x) = x - 1\ g_f(x) = \frac{1}{2}x + \frac{1}{2}$

$$Q^*(s,a) = r(s,a) + \gamma \mathbb{E}_{s'|s,a}\left[V^*(s')\right]$$

$$\pi^*(a|s) = \mu(a|s) \cdot \max\left\{\frac{1}{2} + \frac{Q^*(s,a) - U^*(s)}{2\alpha}, 0\right\}$$

$$V^*(s) = U^*(s) + \alpha \mathbb{E}_{a\sim\mu}\left[\left(\frac{\pi^*(a|s)}{\mu(a|s)}\right)^2\right],$$

$$\mathbb{E}_{a\sim\mu}\left[\max\left\{\frac{1}{2} + \frac{Q^*(s,a) - U^*(s)}{2\alpha}, 0\right\}\right] = 1$$

An approximation $\mathbb{E}_{a\sim\pi^*}\left[\frac{\pi^*(a|s)}{\mu(a|s)}\right] = 1$

$$U^*(s) = V^*(s) - \alpha$$

Lemma 1. *We can get $U^*(s)$ by solving the following optimization problem:*

$$\underbrace{\min_U \mathbb{E}_{a\sim\mu}\left[\mathbb{1}\left(\frac{1}{2} + \frac{Q^*(s,a) - U(s)}{2\alpha} > 0\right)\left(\frac{1}{2} + \frac{Q^*(s,a) - U(s)}{2\alpha}\right)^2\right] + \frac{U(s)}{\alpha}}$$

$$\min_V \mathbb{E}_{\boxed{(s,a)\sim\mathcal{D}}}\left[\mathbb{1}\left(1 + \frac{Q(s,a) - V(s)}{2\alpha} > 0\right)\left(1 + \frac{Q(s,a) - V(s)}{2\alpha}\right)^2 + \frac{V(s)}{\alpha}\right]$$

$$\min_Q \mathbb{E}_{\boxed{(s,a,s')\sim\mathcal{D}}}\left[\left(r(s,a) + \gamma V(s') - Q(s,a)\right)^2\right]$$

$$\max_\pi \mathbb{E}_{\boxed{(s,a)\sim\mathcal{D}}}\left[\mathbb{1}\left(1 + \frac{Q(s,a) - V(s)}{2\alpha} > 0\right)\left(1 + \frac{Q(s,a) - V(s)}{2\alpha}\right)\log\pi(a|s)\right]$$

Method

Exponential Q-learning(EQL) $\alpha \to 0$ Reverse KL divergence, $f(x) = \log(x)$ $g_f(x) = \exp(x-1)$

$$Q^*(s,a) = r(s,a) + \gamma \mathbb{E}_{s'|s,a}\left[V^*(s')\right]$$

$$\pi^*(a|s) = \mu(a|s) \cdot \exp\left(\frac{Q^*(s,a) - U^*(s)}{\alpha} - 1\right)$$

$$V^*(s) = U^*(s) + \alpha \mathbb{E}_{a\sim\mu}\left[\left(\frac{\pi^*(a|s)}{\mu(a|s)}\right)^2 \frac{\mu(a|s)}{\pi^*(a|s)}\right]$$

Without any approximation $U^*(s) = V^*(s) - \alpha$

$$\mathbb{E}_{a\sim\mu}\left[\exp\left(\frac{Q^*(s,a) - V^*(s)}{\alpha}\right)\right] = 1$$

Lemma 2. *We can get $V^*(s)$ by solving the following optimization problem:*

$$\min_V \mathbb{E}_{a\sim\mu}\left[\exp\left(\frac{Q^*(s,a) - V(s)}{\alpha}\right)\right] + \frac{V(s)}{\alpha}$$

$$\min_V \mathbb{E}_{\boxed{(s,a)\sim\mathcal{D}}}\left[\exp\left(\frac{Q(s,a) - V(s)}{\alpha}\right) + \frac{V(s)}{\alpha}\right]$$

$$\min_Q \mathbb{E}_{\boxed{(s,a,s')\sim\mathcal{D}}}\left[\left(r(s,a) + \gamma V(s') - Q(s,a)\right)^2\right]$$

$$\max_\pi \mathbb{E}_{\boxed{(s,a)\sim\mathcal{D}}}\left[\exp\left(\frac{Q(s,a) - V(s)}{\alpha}\right) \log \pi(a|s)\right]$$

Analysis

CQL policy evaluation operator $Q(s,a) = \mathcal{T}^{\pi} Q(s,a) - \beta \left[\frac{\pi(\mathbf{a}|\mathbf{s})}{\mu(\mathbf{a}|\mathbf{s})} - 1 \right]$ The same with SQL

policy improvement $\pi(a|s) \propto \exp(Q(s,a))$

the optimal π is not decided only by Q

IQL policy evaluation

$$\min_{V} \mathbb{E}_{(s,a)\sim\mathcal{D}} \left[\left| \tau - \mathbb{1}\left(Q(s,a) - V(s) < 0\right) \right| \left(Q(s,a) - V(s)\right)^2 \right]$$

$$\min_{Q} \mathbb{E}_{(s,a,s')\sim\mathcal{D}} \left[\left(r(s,a) + \gamma V(s') - Q(s,a) \right)^2 \right],$$

policy extraction

$$\max_{\pi} \mathbb{E}_{(s,a)\sim\mathcal{D}} \left[\exp\left(\frac{Q(s,a) - V(s)}{\alpha} \right) \log \pi(a|s) \right]$$

The same with EQL

the learning objective of V function is not right

In offline RL, policy is not just dependent on Q or V, but still on behavior policy μ

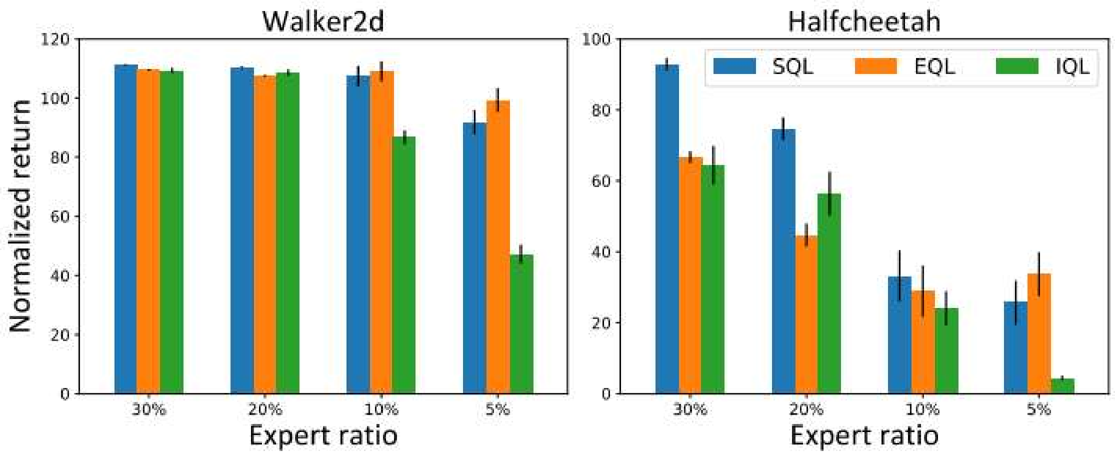
Experiment

ParNeC 模式识别与神经计算研究组 PAttern Recognition and NEural Computing

D4RL benchmark

Dataset	BC	10%BC	BCQ	DT	One-step	TD3+BC	CQL	IQL	SQL	EQL
halfcheetah-m	42.6	42.5	47.0	42.6	48.4	48.3	44.0 ±0.8	47.4 ±0.2	48.3±0.2	47.2±0.3
hopper-m	52.9	56.9	56.7	67.6	59.6	59.3	58.5 ±2.1	66.3 ±5.7	75.5 ±3.4	74.6±2.6
walker2d-m	75.3	75.0	72.6	74.0	81.8	83.7	72.5 ±0.8	72.5 ±8.7	84.2 ±4.6	83.2±4.4
halfcheetah-m-r	36.6	40.6	40.4	36.6	38.1	44.6	45.5 ±0.5	44.2 ±1.2	44.8±0.7	44.5±0.5
hopper-m-r	18.1	75.9	53.3	82.7	97.5	60.9	95.0 ±6.4	95.2 ±8.6	99.7 ±3.3	98.1±3.6
walker2d-m-r	26.0	62.5	52.1	66.6	49.5	81.8	77.2±5.5	76.1 ±7.3	81.2±3.8	76.6±4.2
halfcheetah-m-e	55.2	92.9	89.1	86.8	93.4	90.7	90.7±4.3	86.7±5.3	94.0 ±0.4	90.6±0.5
hopper-m-e	52.5	110.9	81.8	107.6	103.3	98.0	105.4±6.8	101.5 ±7.3	111.8 ±2.2	105.5±2.1
walker2d-m-e	107.5	109.0	109.0	108.1	113.0	110.1	109.6±0.7	110.6±1.0	110.0±0.8	110.2±0.8
antmaze-u	54.6	62.8	78.9	59.2	64.3	78.6	84.8±2.3	85.5 ±1.9	92.2±1.4	93.2 ±2.2
antmaze-u-d	45.6	50.2	55.0	53.0	60.7	71.4	43.4±5.4	66.7 ±4.0	74.0 ±2.3	65.4±2.7
antmaze-m-p	0	5.4	0	0.0	0.3	10.6	65.2±4.8	72.2 ±5.3	80.2 ±3.7	77.5±4.3
antmaze-m-d	0	9.8	0	0.0	0.0	3.0	54.0±11.7	71.0 ±3.2	79.1 ±4.2	70.0±3.7
antmaze-l-p	0	0.0	6.7	0.0	0.0	0.2	38.4±12.3	39.6 ±4.5	53.2 ±4.8	45.6±4.2
antmaze-l-d	0	6.0	2.2	0.0	0.0	0.0	31.6±9.5	47.5 ±4.4	52.3 ±5.2	42.5±4.7
kitchen-c	33.8	-	-	-	-	-	43.8 ±11.2	61.4 ±9.5	76.4 ±8.7	70.3±7.1
kitchen-p	33.9	-	-	-	-	-	49.8±10.1	46.1 ±8.5	72.5±7.4	74.5 ±3.8
kitchen-m	47.5	-	-	-	-	-	51.0±6.5	52.8 ±4.5	67.4 ±5.4	55.6±5.2

Experiment



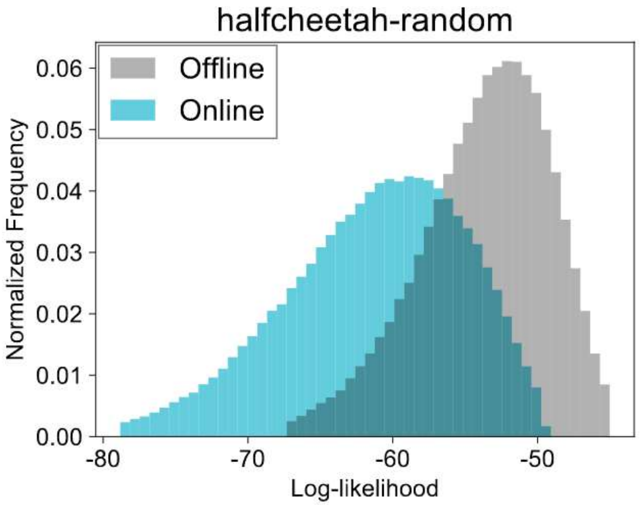
Dataset (AntMaze)		CQL NR	CQL BE	IQL NR	IQL BE	SQL NR	SQL BE
Medium	Vanilla	65.2	13.1	70.0	1.2	75.1	1.6
	Easy	48.2	14.8	49.5	1.1	56.2	1.7
	Medium	14.5	14.7	39.7	1.3	43.3	2.1
	Hard	9.3	64.4	19.6	0.8	24.2	1.9
Large	Vanilla	38.4	13.5	39.6	0.7	50.2	1.4
	Easy	28.1	12.8	35.4	0.7	40.5	1.5
	Medium	6.3	30.6	32.2	0.7	36.7	1.3
	Hard	0	300.5	25.6	0.6	34.2	2.6

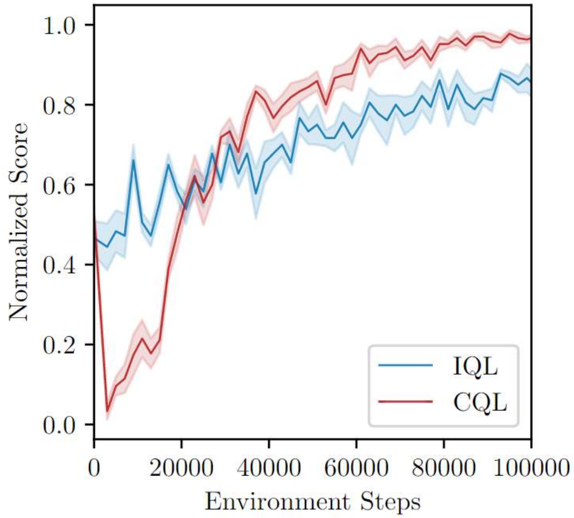
The sparsity term SQL will benefit when the datasets contain a large portion of noisy transitions

In-sample learning brings more robustness than out-of-sample learning

Discussion

The difficulty of offline-to-online





State-action distribution shift, especially in narrow datasets

- underestimate Q of OOD actions excessively MCQ、Cal-QL
- mismatch between actor and critic when offline to online

Discussion

ParNeC 模式识别与神经计算研究组 PAttern Recognition and NEural Computing

mismatch between actor and critic

$$\pi^*(a|s) = \mu(a|s) \cdot \max\left\{g_f\left(\frac{Q^*(s,a) - U^*(s)}{\alpha}\right), 0\right\}$$

SQL $\pi^*(a|s) = \mu(a|s) \cdot \max\left\{1 + \frac{Q^* - V^*}{2\alpha}, 0\right\} \approx \mu(a|s) \cdot \exp\left(\frac{Q^* - V^*}{2\alpha}\right)$ (Taylor expansion of first order)

EQL $\pi^*(a|s) = \mu(a|s) \cdot \exp\left(\frac{Q^* - V^*}{\alpha}\right)$

energy policy in online RL $\pi(a|s) = \exp\left(\frac{Q - V}{\alpha}\right)$

Offline policy $\pi_{\text{offline}}(a|s) \propto \mu(a|s)\pi_{\text{online}}(a|s)$

Offline critic cannot be used for online updates directly!

Discussion

$\pi_{offline}$ is needed, so we should remodel a critic which matches with $\pi_{offline}$ in online RL

policy evaluation of online RL
access to offline dataset
→ off policy evaluation(OPE) of online RL

Extrapolation Error

- Absent Data — **Assumption 1.** *Assume* $\pi(a|s) > 0 \Rightarrow \mu(a|s) > 0$

  $$\mu(a|s) = 0 \Rightarrow \pi(a|s) = 0$$

- Model Bias — add a few online trajectories

- Training Mismatch — $\pi_{\text{offline}}(a|s) = \mu(a|s) \cdot \max\left\{g_f\left(\frac{Q^*(s,a) - u^*(s)}{\alpha}\right), 0\right\}$

  $$\text{IS weight} = \frac{\pi_{offline}(a|s)}{\mu(a|s)}$$ without estimate μ

Discussion

safe offline-to-online

1. obtain $\pi_{offline}$ via offline RL method

2. remodel Q and V by offline dataset or adding some online trajectories

3. online safe exploration

Thanks