南京航空航天大學
Nanjing University of Aeronautics and Astronautics

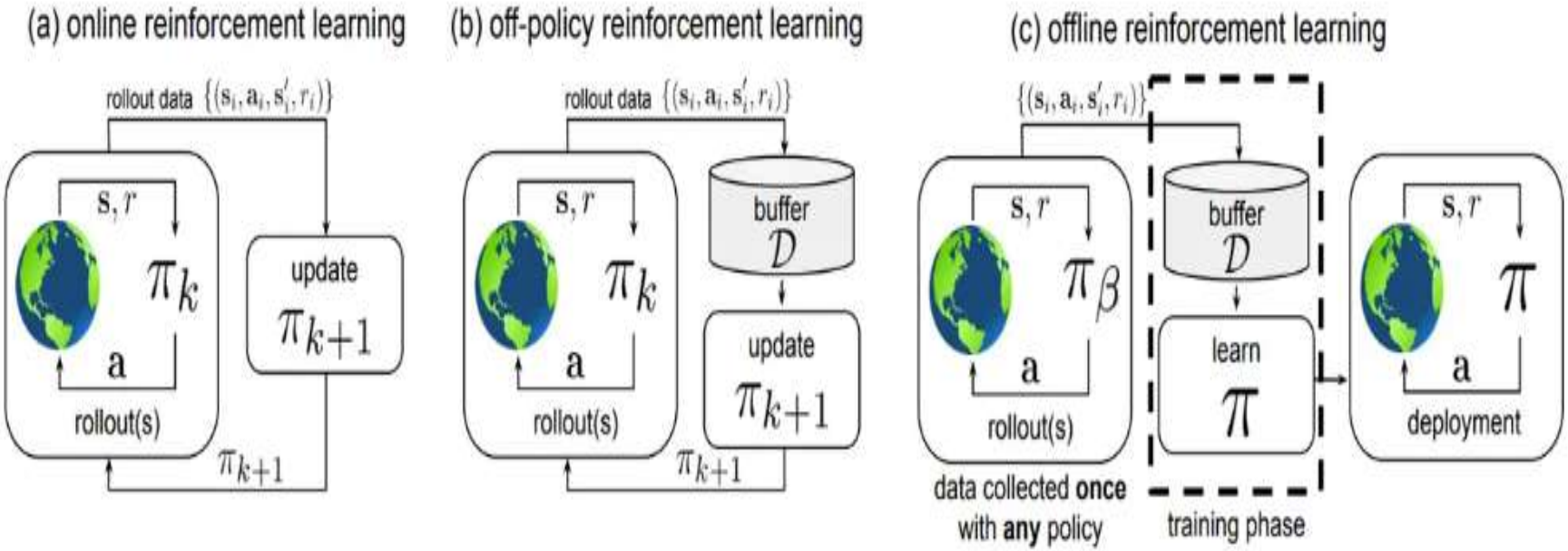
模式分析与机器智能
工业和信息化部重点实验室
MIIT Key Laboratory of
Pattern Analysis & Machine Intelligence

DUAL RL: UNIFICATION AND NEW METHODS FOR REINFORCEMENT AND IMITATION LEARNING

ICLR | 2024

Background

- **Reinfocement learning**



$$V^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s\right]$$

$$\mathcal{T}_r^{\pi} Q(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim p(\cdot|s,a), a' \sim \pi(\cdot|s')}[Q(s',a')].$$

$$\mathcal{T}_r V(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim p(\cdot|s,a)}[V(s')].$$

Background

- **Limitation**
  - offline algorithms often face significant challenges when learning from off-policy data
    - $\begin{cases} \text{instability} \\ \text{Overestimation} \end{cases}$ （**ood**）

- **Offline Reinfocement learning**

$$J(\pi) = \mathbb{E}_{\tau \sim p_\pi(\tau)} \left[\sum_{t=0}^{H} \gamma^t r(\mathbf{s}_t, \mathbf{a}_t) \right] . \xrightarrow{\text{constraint}} \max_{\pi_\theta} J(\pi_\theta) - \alpha \mathbb{E}_{(s,a) \sim \mathcal{D}} \left[\mathcal{C}(s,a) \right],$$

often require that the majority of training data be nearly on-policy to achieve high learning performance （suboptimal）

Background

- **Motivation**
  - **Overcoming the biases and instability**
  - **Reducing the reliance on nearly on-policy data**
  - **Establishing a unified theoretical framework**

- **Utilizing Convex Programs for Policy Optimization(DICE**（DIstribution Correction Estimation)）
  - Formulating RL as a Convex Program
  - Conversion Using Lagrangian Duality

Priliminary

1. 凸集

1.1 定义：

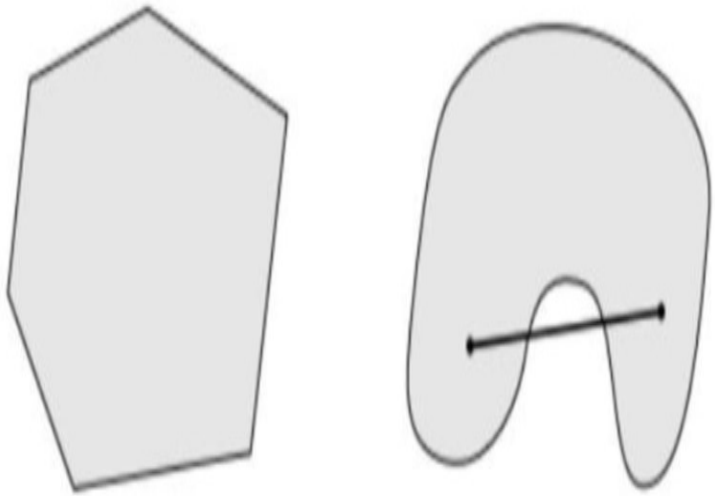
C 是凸集，如果对于任意的 $x, y \in C$ 和任意的 $\theta \in \mathbb{R}$ 满足 $0 \leq \theta \leq 1$ 时，
$\theta x + (1-\theta)y \in C$ 恒成立

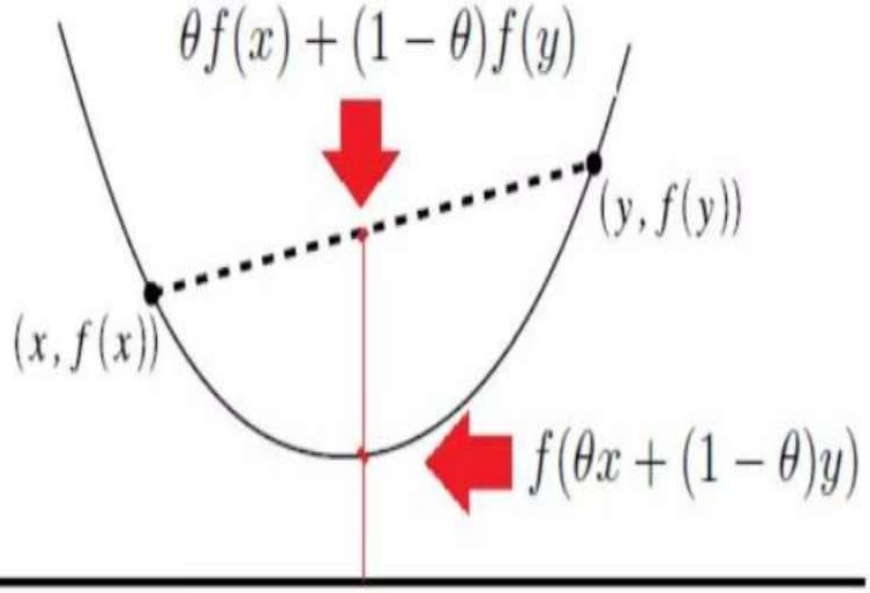
2. 凸函数

2.1定义：

定义在 $\mathbb{R}^n \to \mathbb{R}$ 上的函数 f 是凸函数，如果它的定义域 $\mathbb{D}(f)$ 是一个凸集且对任意的 $x, y \in \mathbb{D}$
和 $0 \leq \theta \leq 1$，$f(\theta x + (1-\theta y)) \leq \theta f(x) + (1-\theta) f(y)$ 恒成立

Priliminary





Priliminary

- **对偶**
  - 如果你用不超过100块钱买水果，
    - 问题1： 买到单价最便宜的水果
    - 问题2： 买到重量最多的水果

原始问题 (Primal Problem):

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad g_i(x) \leq 0, \quad i = 1, \ldots, m \\ &\qquad\quad h_j(x) = 0, \quad j = 1, \ldots, p \end{aligned}$$

对偶问题(dual problem) ：

$$\begin{aligned} &\text{maximize} \quad d(\lambda, \nu) \\ &\text{subject to} \quad \lambda_i \geq 0, \quad i = 1, \ldots, m \end{aligned}$$

其中 $d(\lambda,\nu) = \inf_x L(x,\lambda,\nu)$ ， $L(x,\lambda,\nu) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{j=1}^{p} \nu_j h_j(x)$

Priliminary

原问题的**拉格朗日对偶(Dual)问题**, 记为 D, 定义如下:

$$(\mathrm{D.}) \qquad \max_{\lambda,v} \quad d(\lambda, v) = min_x L(x, \lambda, v)$$
$$\text{s.t.} \qquad \lambda \geq 0$$

由于对偶函数是凹函数，故拉格朗日对偶问题一定是凸优化问题，其对应的最优解为 λ^*, v^* (最优拉格朗日乘子)，对应的最优值为 d^*, **有**$d^* \leq p^*$.

目的是求原始问题的最优值。得一个简单的下界函数, 然后通过最大化这个下界函数, 来逼近原问题的最优解.

Priliminary

- **REINFORCEMENT LEARNING VIA LAGRANGIAN DUALITY**

Consider the regularized policy learning problem

$$\max_{\pi} J(\pi) = \mathbb{E}_{d^{\pi}(s,a)}[r(s,a)] - \alpha D_f(d^{\pi}(s,a) \,||\, d^{O}(s,a)), \tag{1}$$

where $D_f(d^{\pi}(s,a) \,||\, d^{O}(s,a))$ is a conservatism regularizer that encourages the visitation distribution of π to stay close to some distribution d^{O}, and α is a temperature parameter that balances the expected return and the conservatism.

rewritten as a convex problem

primal-Q

$$\max_{\pi} J(\pi) = \max_{\pi} \Big[\max_{d} \mathbb{E}_{d(s,a)}[r(s,a)] - \alpha D_f(d(s,a) \,||\, d^{O}(s,a)) \tag{2}$$
$$\text{s.t } d(s,a) = (1-\gamma) d_0(s).\pi(a|s) + \gamma \textstyle\sum_{s',a'} d(s',a') p(s|s',a') \pi(a|s), \ \forall s \in \mathcal{S}, a \in \mathcal{A} \Big].$$

dual-Q

$$\max_{\pi} \min_{Q} (1-\gamma) \mathbb{E}_{s \sim d_0, a \sim \pi(s)}[Q(s,a)] + \alpha \mathbb{E}_{(s,a) \sim d^{O}}[f^{*}([\mathcal{T}_r^{\pi} Q(s,a) - Q(s,a)]/\alpha)], \tag{3}$$

where f^{*} is the convex conjugate of f.

Priliminary

Directly use dual-Q causes overconstraint

primal-V

$$\max_{d \geqslant 0} \mathbb{E}_{d(s,a)}[r(s,a)] - \alpha D_f(d(s,a) \,\|\, d^O(s,a))$$
$$\text{s.t} \sum_{a \in \mathcal{A}} d(s,a) \models (1-\gamma) d_0(s) + \gamma \sum_{(s',a') \in \mathcal{S} \times \mathcal{A}} d(s',a') p(s|s',a'), \ \forall s \in \mathcal{S}. \tag{4}$$

dual-V

$$\min_V (1-\gamma)\mathbb{E}_{s \sim d_0}[V(s)] + \alpha \mathbb{E}_{(s,a) \sim d^O}\left[f_p^*\left([\mathcal{T}V(s,a) - V(s)]/\alpha\right)\right], \tag{5}$$

求对偶V等价于求原始问题V，并且对偶V不需要对数据进行约束, DUAL-V作为辅助优化目标对s，a进行优化选择

Priliminary

- **A UNIFIED PERSPECTIVE ON RL AND IL THROUGH DUALITY**

Imitation Learning

dual-Q $\max_{\pi} \min_{Q}(1-\gamma)\mathbb{E}_{d_0(s),\pi(a|s)}[Q(s,a)] + \alpha\mathbb{E}_{s,a\sim d^E}[f^*([\mathcal{T}_0^{\pi}Q(s,a) - Q(s,a)]/\alpha)]. \quad (6)$

Reinforcment Learning

dual-Q $\max_{\pi(a|s)} \min_{Q(s,a)}(1-\gamma)\mathbb{E}_{\rho_0(s),\pi(a|s)}[Q(s,a)] + \mathbb{E}_{s,a\sim d^S}\left[f^*(\mathcal{T}_{rimit}^{\pi}Q(s,a) - Q(s,a))\right], \quad (7)$

Methodology: RECOIL

- **RElaxed Coverage for Off-policy Imitation Learning (ReCOIL),**

mixture distributions $d^S_{\text{mix}} := \beta d(s,a) + (1-\beta) d^S(s,a)$ and $d^{E,S}_{\text{mix}} := \beta d^E(s,a) + (1-\beta) d^S(s,a)$,

$$\texttt{primal-Q} \qquad \max_{d(s,a)} -D_f(d^S_{\text{mix}}(s,a) \,||\, d^{E,S}_{\text{mix}}(s,a))$$

$$\text{s.t}\ \forall s \in \mathcal{S}, a \in \mathcal{A},\ \ d(s,a) = (1-\gamma) d_0(s) \pi(a|s) + \gamma \textstyle\sum_{(s',a') \in \mathcal{S} \times \mathcal{A}} d(s',a') p(s|s',a') \pi(a|s). \quad (8)$$

Theorem 1. *(ReCOIL objective) The* `dual-Q` *problem to the mixture distribution matching objective in Eq. 8 is given by:*

$$\max_\pi \min_Q \beta(1-\gamma)\mathbb{E}_{d_0,\pi}[Q(s,a)] + \mathbb{E}_{s,a \sim d^{E,S}_{mix}}[f^*(\mathcal{T}^\pi_0 Q(s,a) - Q(s,a))] - (1-\beta)\mathbb{E}_{s,a \sim d^S}[\mathcal{T}^\pi_0 Q(s,a) - Q(s,a)] \quad (9)$$

Methodology: RECOIL

Algorithm 1: ReCOIL (offline, χ^2)

1: Initialize Q_ϕ, V_θ, and π_ψ, mixing ratio β, conservatism τ, temperature α
2: $\mathcal{D}^{\mathcal{S}} = (s, a, s')$ be suboptimal dataset
3: $\mathcal{D}^{\mathcal{E}} = (s, a, s')$ be expert dataset.
4: **for** $t = 1..T$ iterations **do**
5: Train Q_ϕ using $\min_\phi \mathcal{L}(\phi)$:
6: Train V_θ using $\min_\theta \mathcal{J}(\theta)$
7: Update π_ψ via $\max_\psi \mathcal{M}(\psi)$:
8: **end for**

$$\mathcal{L}(\phi) = \beta(\mathbb{E}_{d^S,\pi(a|s)}[Q_\phi(s,a)] - \mathbb{E}_{d^E(s,a)}[Q_\phi(s,a)]) + 0.25\,\mathbb{E}_{s,a\sim d^{E,S}_{\text{mix}}(s,a)}[(\gamma V_\theta(s') - Q_\phi(s,a))^2]. \quad (11)$$

$$\mathcal{J}(\theta) = \mathbb{E}_{s,a\sim d^{E,S}(s,a)}[\exp((Q_\phi(s,a) - V_\theta(s))/\tau) + (Q_\phi(s,a) - V_\theta(s))/\tau)]. \quad (12)$$

$$\mathcal{M}(\psi) = \max_\psi \mathbb{E}_{s,a\sim d^{E,S}_{\text{mix}}(s,a)}[\exp(\bar{\alpha}(Q_\phi(s,a) - V_\theta(s)))\log(\pi_\psi(a|s))]. \quad (13)$$

Methodology: f-DVL

Consider a rewriting of dual-V with the temperature parameter λ and a chosen surrogate function $\bar{f}_p^*$ that extends the domain of f_p^* to $\mathbb{R}$. We discuss the need for a surrogate function below.

$$\min_V (1-\lambda)\mathbb{E}_{s\sim d^O}[V(s)] + \lambda\mathbb{E}_{(s,a)\sim d^O}\left[\bar{f}_p^*\left(\bar{Q}(s,a) - V(s)\right)\right], \tag{14}$$

where $\bar{Q}(s,a)$ denotes $\texttt{stop-gradient}(r(s,a) + \gamma\sum_{s'} p(s'|s,a)V(s'))$. Let x be a random variable of distribution D. Eq (14) can be considered as a special instance of the following problem:

$$\min_v (1-\lambda)v + \lambda\mathbb{E}_{x\sim D}\left[\bar{f}_p^*\left(x - v\right)\right], \tag{15}$$

Algorithm 2: f-DVL (Under Stochastic Dynamics)

1: Initialize Q_ϕ, V_θ, π_ψ, temperature α, weight λ
2: Let $\mathcal{D} = (s, a, r, s')$ be offline dataset
3: **for** $t = 1..T$ iterations **do**
4: Train Q_ϕ by minimizing: $\mathbb{E}_{s,a,s'\sim\mathcal{D}}\left[\left(Q_\phi(s,a) - (r(s,a) + \gamma V_\theta(s'))\right)^2\right]$.
5: Train V_θ by minimizing Eq 14 with surrogate $\bar{f}_p^*$
6: Update π_ψ by maximizing: $\mathbb{E}_{s,a\sim\mathcal{D}}\left[e^{\alpha(Q_\phi(s,a) - V_\theta(s))}\log\pi_\psi(s|a)\right]$.
7: **end for**

Experiments

Suboptimal Dataset	Env	RCE	ORIL	SMODICE	BC (only expert data)	BC (full dataset)	IQ-Learn (offline)	ReCOIL	Expert
random+ expert	hopper	51.41±38.63	73.93±11.06	**101.61±7.69**	4.52±1.42	5.64±4.83	1.85 ±2.19	**108.18±3.28**	111.33
	halfcheetah	64.19±11.06	60.49±3.53	**80.16±7.30**	2.2±0.01	2.25±0.00	4.83±7.99	**80.20±6.61**	88.83
	walker2d	20.90±26.80	2.86±3.39	**105.86±3.47**	0.86±0.61	0.91±0.5	0.57±0.09	**102.16±7.19**	106.92
	ant	105.38±14.15	73.67±12.69	**126.78±5.12**	5.17±5.43	30.66±1.35	42.23±20.05	**126.74±4.63**	130.75
random+ few-expert	hopper	25.31±18.97	42.04±13.76	60.11±18.28	4.84±3.83	3.0±0.54	1.37 ±1.23	**97.85±17.89**	111.33
	halfcheetah	2.99±1.07	2.84±5.52	2.28±0.62	-0.93±0.35	2.24±0.01	1.14±1.94	**76.92±7.53**	88.83
	walker2d	40.49±26.52	3.22±3.29	**107.18±1.87**	0.98±0.83	0.74±0.20	0.39±0.27	83.23±19.00	106.92
	ant	**67.62±15.81**	25.41 ± 8.58	-6.10±7.85	0.91±3.93	35.38±2.66	32.99±3.12	**67.14± 8.30**	130.75
medium+ expert	hopper	58.71±34.06	61.68±7.61	49.74±3.62	16.09±12.80	59.25±3.71	12.90±24.00	**88.51±16.73**	111.33
	halfcheetah	65.14±13.82	54.66±0.88	59.50±0.82	-1.79±0.22	42.45± 0.42	25.67±20.82	**81.15±2.84**	88.83
	walker2d	**96.24±14.04**	8.19±7.70	2.62±0.93	2.43±1.82	72.76±3.82	59.37±30.14	**108.54±1.81**	106.92
	ant	**86.14±38.59**	102.74±6.63	104.95±6.43	0.86±7.42	95.47±10.37	37.17±41.15	**120.36±7.67**	130.75
medium few-expert	hopper	**66.15±35.16**	17.40±15.15	47.61±7.08	7.37±1.13	46.87±5.31	11.05±20.59	**50.01±10.36**	111.33
	halfcheetah	**61.14±18.31**	43.24±0.75	46.45±3.12	-1.15±0.06	42.21±0.06	26.27±20.24	**75.96±4.54**	88.83
	walker2d	**85.28±34.90**	6.81±6.76	6.00±6.69	2.02±0.72	70.42±2.86	73.30±2.85	**91.25±17.63**	106.92
	ant	**67.95±36.78**	81.53±8.618	81.53±8.618	-10.45±1.63	81.63±6.67	35.12±50.56	**110.38±10.96**	130.75
cloned+expert	pen	19.60±11.40	-3.10±0.40	-3.36±0.71	13.95±11.04	34.94±11.10	2.18±8.75	**95.04±4.48**	106.42
	door	0.08± 0.15	-0.33±0.01	0.25± 0.54	-0.22±0.05	0.011±0.00	0.07±0.02	**102.75±4.05**	103.94
	hammer	1.95±3.89	0.25± 0.01	0.15± 0.078	2.41±4.48	5.45± 7.84	0.27±0.02	**95.77±17.90**	125.71
	relocate	-0.25±0.04	-0.29±0.01	1.75±3.85	-0.17±0.04	-0.24± 0.01	-0.1±0.12	**67.43±14.60**	118.39
human+expert	pen	17.81±5.91	-3.38±2.29	-2.20±2.40	13.83±10.76	90.76±25.09	14.29±28.82	**103.72±2.90**	106.42
	door	-0.05±0.05	-0.33±0.01	-0.20± 0.11	-0.03±0.05	103.71±1.22	5.6±7.29	**104.70±0.55**	103.94
	hammer	5.00±5.64	1.89±0.70	-0.07±0.39	0.18±0.14	122.61±4.85	5.32±1.38	**125.19±3.29**	125.71
	relocate	0.02±0.10	-0.29±0.01	-0.16±0.04	-0.13±0.11	81.19±7.73	-0.04±0.22	**91.98± 2.89**	118.39
partial+expert	kitchen	6.875±9.24	0.00±0.00	39.16± 1.17	2.5±5.0	45.5±1.87	0.0±0.0	**60.0±5.70**	75.0
mixed+expert	kitchen	1.66±2.35	0.00±0.00	42.5±2.04	2.2±3.8	42.1±1.12	0.0±0.0	**52.0±1.0**	75.0

Table 2: The normalized return obtained by different offline IL methods trained on the D4RL suboptimal datasets with 1 expert trajectory. Methods with avg. perf within the std-dev of the top performing method is highlighted.

Experiments

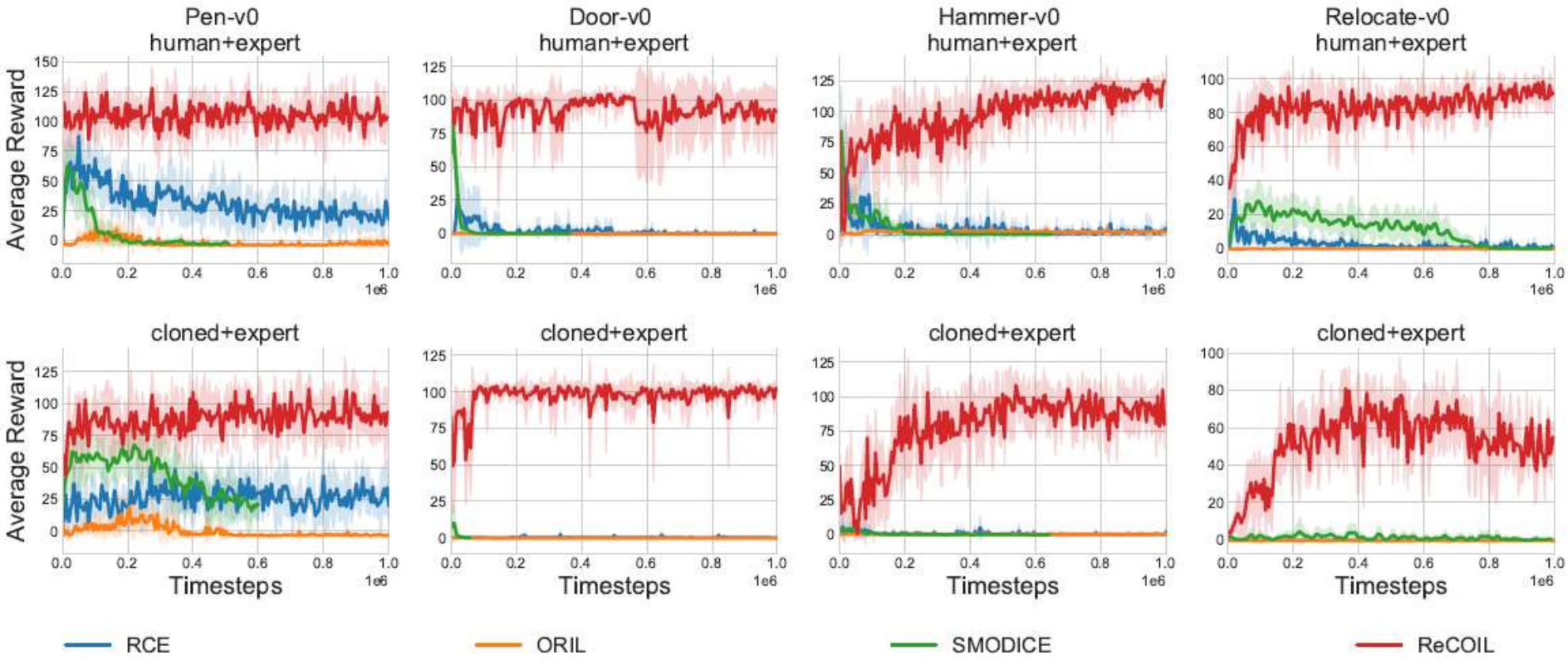


Figure 10: Learning curves for ReCOIL showing that it outperforms baselines in the setting of learning to imitate from diverse offline data. The results are averaged over 7 seeds

Experiments

Dataset	BC	10%BC	DT	TD3+BC	CQL	IQL	XQL(r)	f-DVL (χ^2)	f-DVL (TV)
halfcheetah-medium-v2	42.6	42.5	42.6	**48.3**	44.0	**47.4**	47.4	**47.7**	47.5
hopper-medium-v2	52.9	56.9	67.6	59.3	58.5	66.3	**68.5**	63.0	64.1
walker2d-medium-v2	75.3	75.0	74.0	**83.7**	72.5	78.3	81.4	80.0	81.5
halfcheetah-medium-replay-v2	36.6	40.6	36.6	**44.6**	**45.5**	44.2	44.1	42.9	**44.7**
hopper-medium-replay-v2	18.1	75.9	82.7	60.9	95.0	94.7	95.1	90.7	**98.0**
walker2d-medium-replay-v2	26.0	62.5	66.6	**81.8**	77.2	73.9	58.0	52.1	68.7
halfcheetah-medium-expert-v2	55.2	**92.9**	86.8	90.7	91.6	86.7	90.8	89.3	91.2
hopper-medium-expert-v2	52.5	**110.9**	107.6	98.0	105.4	91.5	94.0	105.8	93.3
walker2d-medium-expert-v2	107.5	109.0	108.1	**110.1**	108.8	**109.6**	**110.1**	**110.1**	**109.6**
antmaze-umaze-v0	54.6	62.8	59.2	78.6	74.0	**87.5**	47.7	83.7	**87.7**
antmaze-umaze-diverse-v0	45.6	50.2	53.0	71.4	**84.0**	62.2	51.7	50.4	48.4
antmaze-medium-play-v0	0.0	5.4	0.0	10.6	61.2	**71.2**	31.2	56.7	**71.0**
antmaze-medium-diverse-v0	0.0	9.8	0.0	3.0	53.7	**70.0**	0.0	48.2	60.2
antmaze-large-play-v0	0.0	0.0	0.0	0.2	15.8	39.6	10.7	36.0	**41.7**
antmaze-large-diverse-v0	0.0	6.0	0.0	0.0	14.9	**47.5**	31.28	44.5	39.3
kitchen-complete-v0	65.0	-	-	-	43.8	62.5	56.7	**67.5**	61.3
kitchen-partial-v0	38.0	-	-	-	49.8	46.3	48.6	58.8	**70.0**
kitchen-mixed-v0	51.5	-	-	-	51.0	51.0	40.4	**53.75**	52.5

Table 3: The normalized return of offline RL methods on D4RL tasks. XQL(r) denotes the results obtained under the standard evaluation protocol. Results aggregated over 7 seeds. Highlighted results are within one performance point of the best-performing algorithm.