NEURAL INFORMATION
PROCESSING SYSTEMS

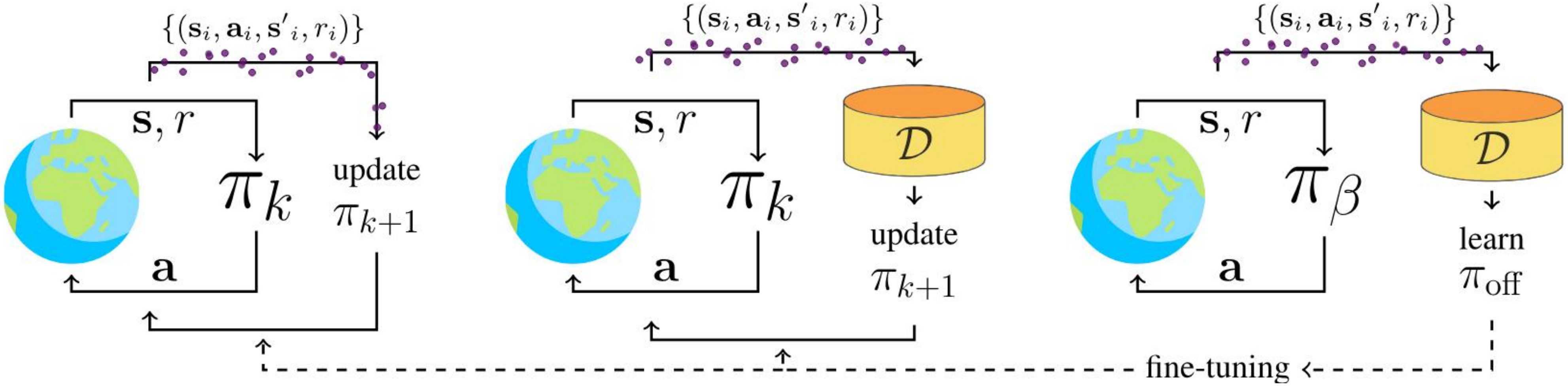
Optimistic Critic Reconstruction and Constrained Fine-Tuning for General Offline-to-Online RL

Qin-Wen Luo*[1] , Ming-Kun Xie*[1,2], Ye-Wen Wang*[1] , Sheng-Jun Huang✉[1]

[1] Nanjing University of Aeronautics and Astronautics, Nanjing, China

[2] RIKEN Center for Advanced Intelligence Project, Tokyo, Japan

Offline-to-online RL



on-policy

off-policy

offline

combat distribution shift

policy constraint/value regularization

maintain pessimism

hinder efficiency/individually designed

Two mismatches

online

$$J(\pi) = \mathbb{E}_{\pi}\left[\sum\nolimits_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right]$$

$$\begin{cases} Q^{k+1}(s,a) = r(s,a) + \mathbb{E}_{s' \sim P(\cdot|s,a), a' \sim \pi_k(\cdot|s')}\left[Q^k(s',a')\right] \\ \pi_{k+1} = \underset{\pi}{\arg\max}\; Q^{\pi_k}(s,a) \end{cases}$$

policy evaluation

policy improvement

offline

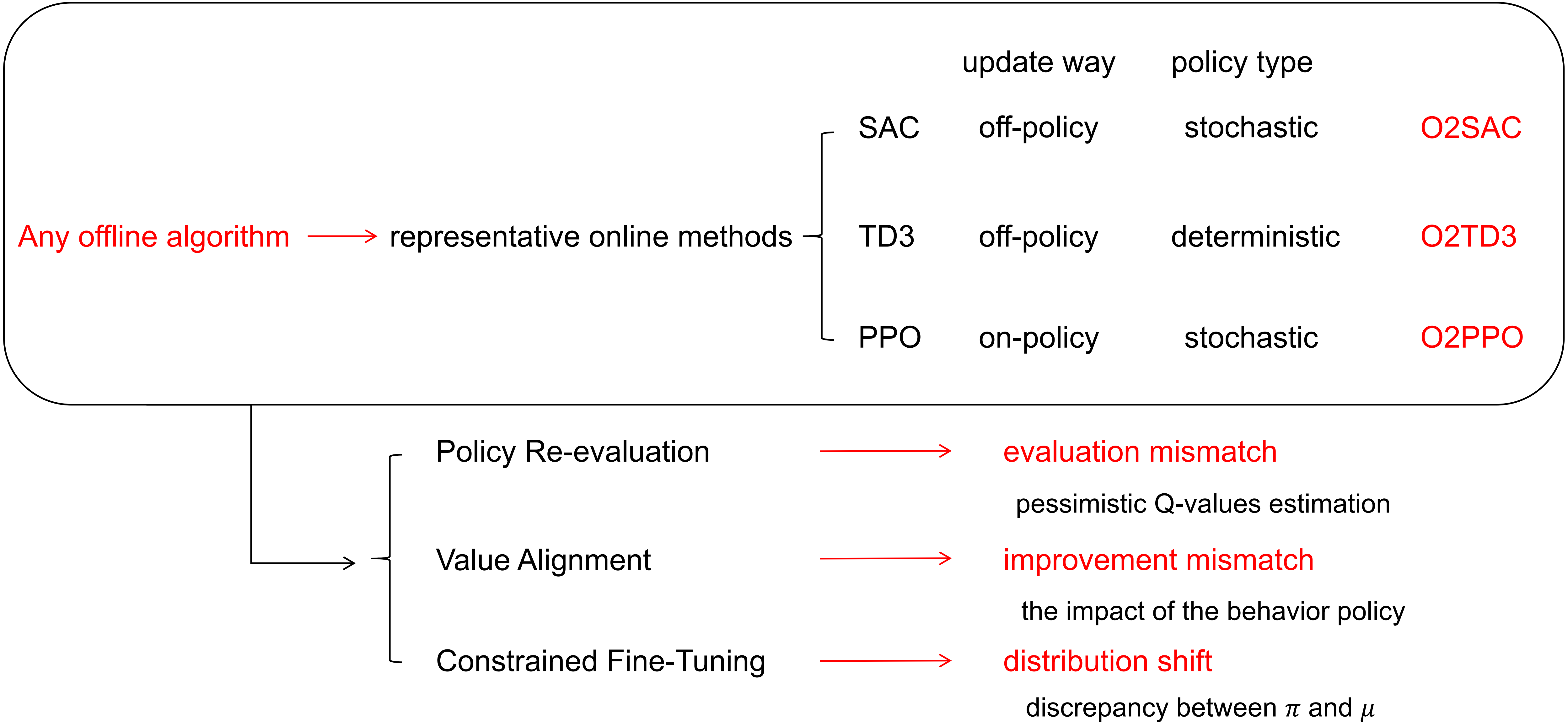
$$J(\pi) = \mathbb{E}_{\pi}\left[\sum\nolimits_{t=0}^{\infty} \gamma^t \left(r(s_t, a_t) - \alpha \cdot f\left(\frac{\pi(a_t|s_t)}{\mu(a_t|s_t)}\right)\right)\right]$$

evaluation mismatch

improvement mismatch

$$\begin{cases} Q^{k+1}(s,a) = r(s,a) + \mathbb{E}_{s' \sim P(\cdot|s,a), a' \sim \pi_k(\cdot|s')}\left[Q^k(s',a') - \alpha \cdot f\left(\frac{\pi_k(a'|s')}{\mu(a'|s')}\right)\right] \\ \pi_{k+1} = \underset{\pi}{\arg\max}\; Q^{\pi_k}(s,a) - \alpha \cdot f\left(\frac{\pi_k(a|s)}{\mu(a|s)}\right) \end{cases}$$

Method



Any offline algorithm → representative online methods

	update way	policy type	
SAC	off-policy	stochastic	O2SAC
TD3	off-policy	deterministic	O2TD3
PPO	on-policy	stochastic	O2PPO

Policy Re-evaluation → evaluation mismatch

pessimistic Q-values estimation

Value Alignment → improvement mismatch

the impact of the behavior policy

Constrained Fine-Tuning → distribution shift

discrepancy between π and μ

Policy Re-evaluation

off-policy evaluation （OPE）

- Fitted Q-learning (FQE) → O2SAC, O2TD3 consistent with online update

  theoretical guarantee of FQE

  $$\max_{(s,a)\in S\times A}\frac{d^{\pi_\theta}(s,a)}{d^{\mu}(s,a)}\leq C.$$

  $$|Q^\pi-\hat{Q}^\pi|\leq\frac{1-\gamma^K}{1-\gamma}\sqrt{C\epsilon}+\gamma^K\bar{V},\ \textit{where}\ \epsilon:=\frac{22\bar{V}^2\log(|\mathcal{F}|/\delta)}{|\mathcal{D}|}+20d_F^\pi.$$

- Fitted returns → O2PPO the fitted retrurns only approximate $V^\mu(s)$

↓

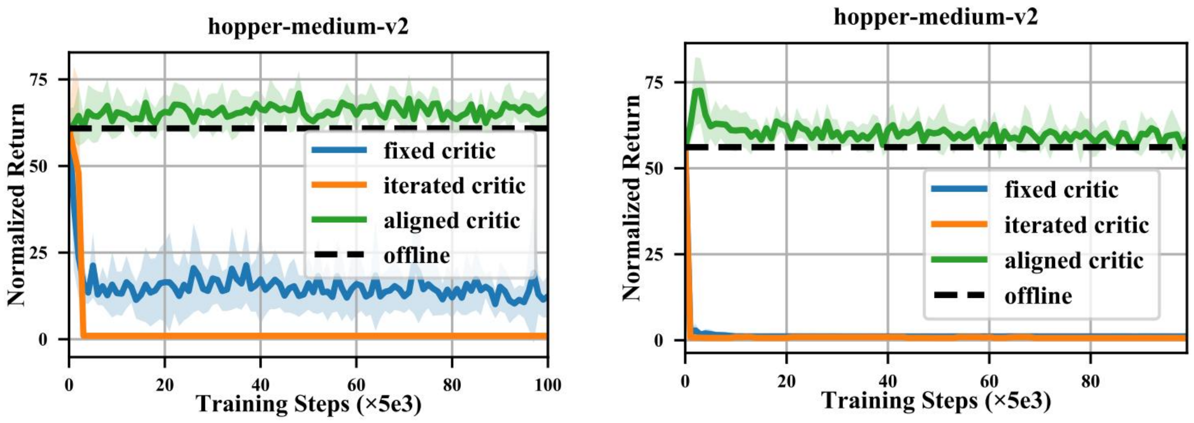
optimistic critic

Value Alignment

inconsistent offline update

imperfect data for FQE

improvement mismatch



(a) The performance of SAC

(b) The performance of TD3

O2SAC: $Q(s,a)=V(s)+\alpha\log\pi\left(a|s\right) \longrightarrow Q'_{\mu}(s,a)=\min\left(Q_{\bar{\mu}}(s,\dot{a})-\alpha\left(\log\pi_{\text{off}}\left(\dot{a}|s\right)-\log\pi_{\text{off}}\left(a|s\right)\right),Q_{\bar{\mu}}(s,a)\right)$

value bound $V_{fqe}(s)\leq V_{align}(s)\leq V_{\dot{a}}(s)$

O2TD3: $Q(s,a)/Q(s,\dot{a})\sim N(\dot{a},\Sigma) \longrightarrow Q'(s,a)=\min\left(Q_{\bar{\mu}}(s,\tilde{a}),\frac{Q(s,\dot{a})}{1+k\cdot\max\left(d(a,\dot{a})^{2},\sigma^{2}\right)}\right)$

O2PPO: $A_{\alpha}(s,a)=\alpha\log\pi_{\text{off}}(a|s)+\alpha\mathcal{H}(\pi_{\text{off}}(\cdot|s)) \longrightarrow A'(s,a)=A(s,a)+\beta A_{\alpha}(s,a)$

equivalent constraint $-\underset{s\sim R}{\mathbb{E}}\ \underset{a\sim\pi_{\theta_k}(\cdot|s)}{\mathbb{E}}[\frac{\pi_{\theta}}{\pi_{\theta_k}}A_{\alpha}(s,a)]\Longleftrightarrow CELoss(\pi_{\theta},\pi_{\text{ref}})+C$

Constrained Fine-Tuning

Constrained MDP

$$\max \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma_t r_t(s_t, a_t)] \quad \text{s.t. } \mathbb{E}_\pi[f(\pi(a_t|s_t), \pi_{\text{ref}}(a_t|s_t))] < \tau$$

$$\Downarrow$$

$$\begin{aligned}
L(\theta) &= \max \mathbb{E}_{\pi_\theta}[Q_\mu^{\pi_\theta}(s,a) - \lambda f(\pi_\theta(a|s), \pi_{\text{ref}}(a|s))] \\
L(\mu) &= \min \mathbb{E}_{(s,a,r,s')\sim R}[(Q_\mu^{\pi_\theta}(s,a) - y)^2] \\
y &= r + \gamma \mathbb{E}_{a' \sim \pi_\theta(\cdot|s')}[Q_{\bar{\mu}}^{\pi_\theta}(s',a') - \lambda f(\pi_\theta(a'|s'), \pi_{\text{ref}}(a'|s'))] \\
L(\lambda) &= \min_{\lambda \geq 0} -\lambda\,[\mathbb{E}_{\pi_\theta}(f(\pi_\theta(a|s), \pi_{\text{ref}}(a|s))) - \tau]
\end{aligned} \tag{20}$$

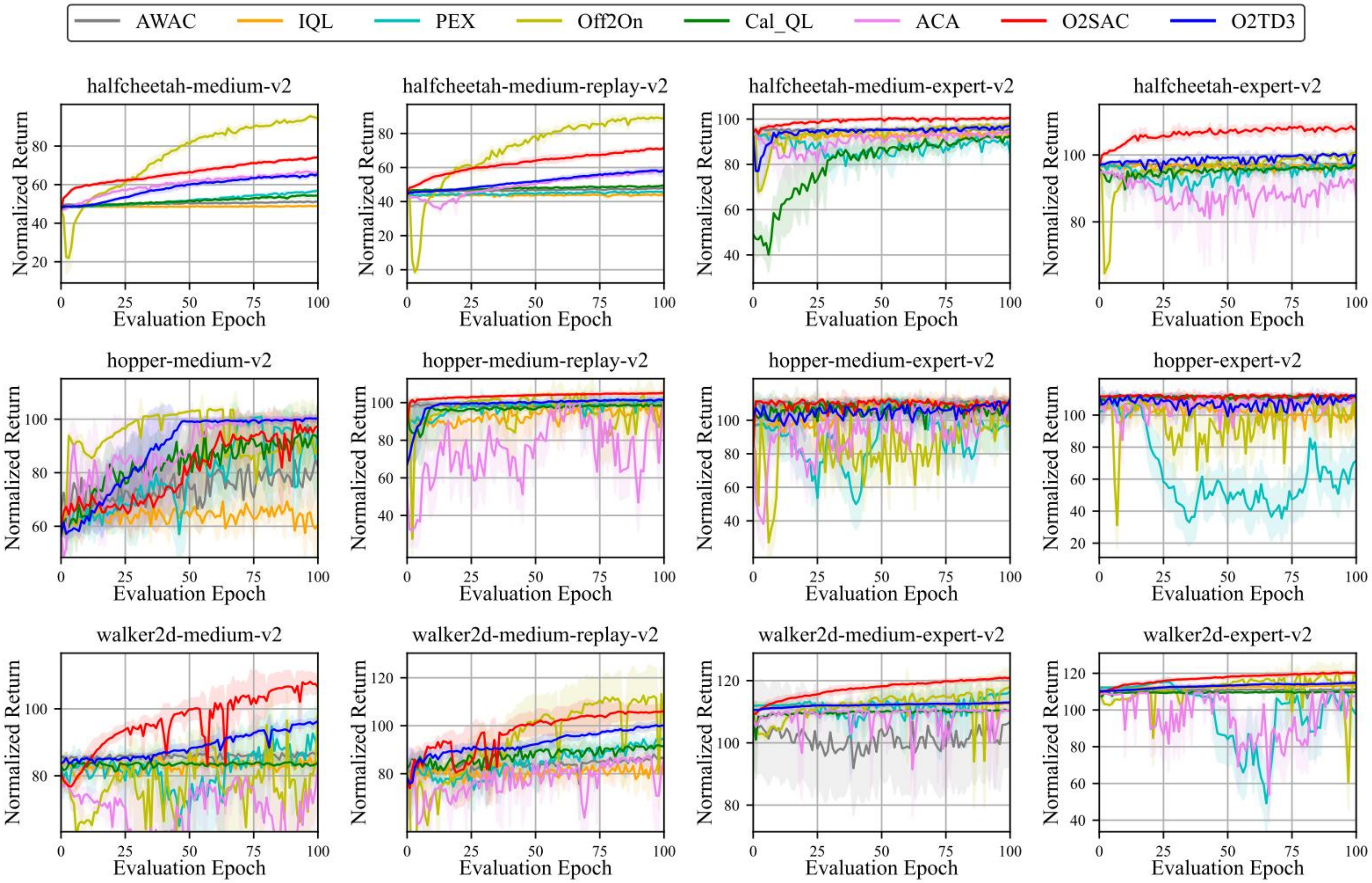
Corollary 4.5. *With the penalty $f(\pi, \pi_{ref})$ defined before and appropriate learning rates, algorithm of Eq. ($\boxed{20}$) almost surely to a fixed point $(\theta^*, \mu^*, \lambda^*)$, where $\lambda^* = 0$, θ^* and μ^* are corresponded to π^* and Q^*, which are optimal in the MDP without constraint.*

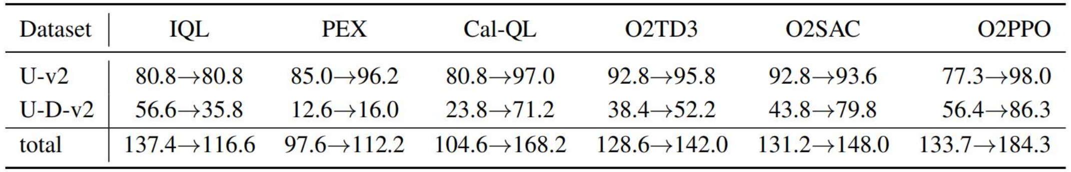
O2SAC: $f(\mu, \phi) = D_{KL}(\mu|\phi)$

O2TD3: $f(\mu, \phi) = MSE(\mu, \phi)$

O2PPO: $A_\alpha(s,a) = \alpha \log \pi_{\text{ref}}(a|s) + \alpha \mathcal{H}(\pi_{\text{ref}}(\cdot|s))$

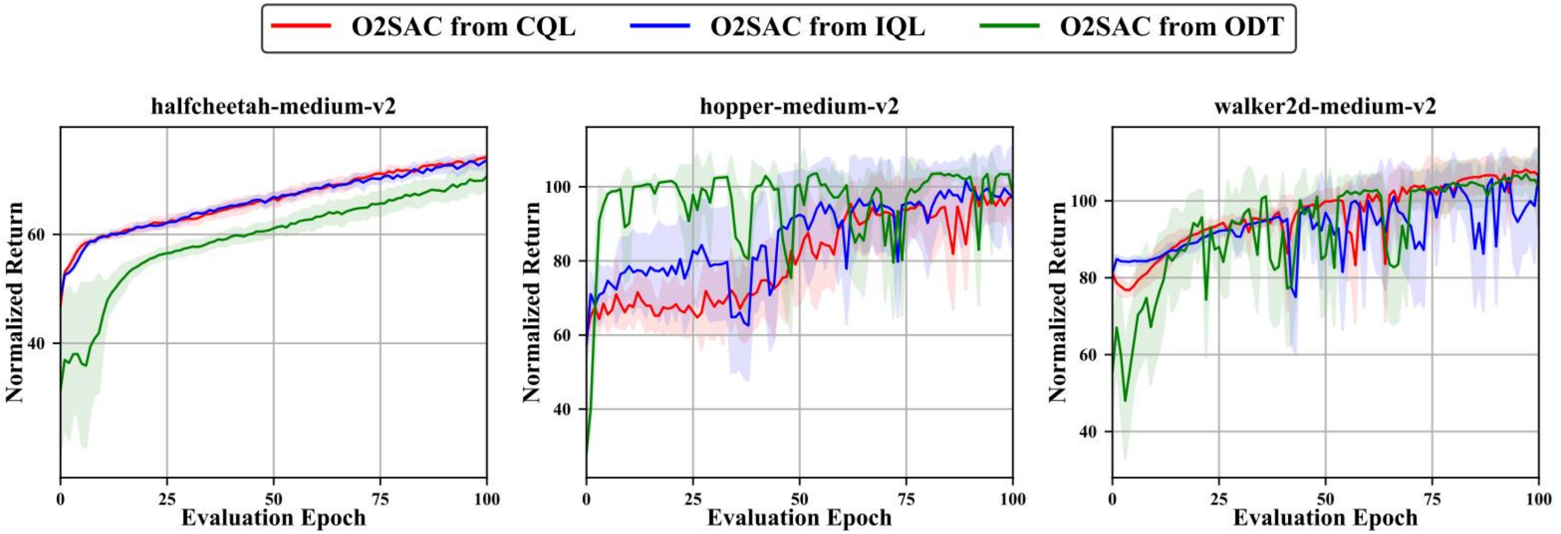
Experiments



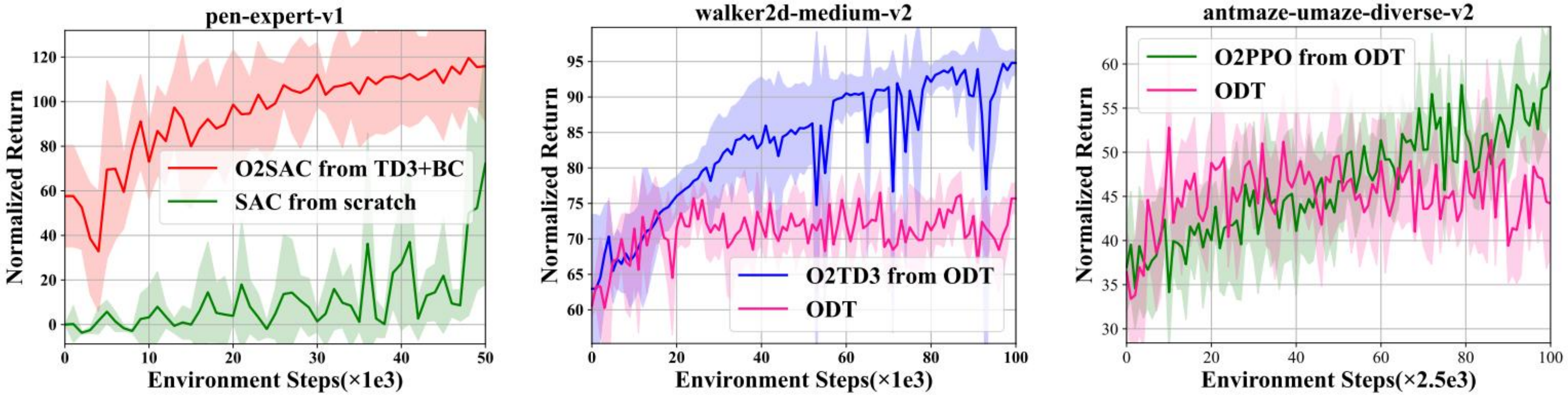


Dataset	IQL	PEX	Cal-QL	O2TD3	O2SAC	O2PPO
U-v2	80.8→80.8	85.0→96.2	80.8→97.0	92.8→95.8	92.8→93.6	77.3→98.0
U-D-v2	56.6→35.8	12.6→16.0	23.8→71.2	38.4→52.2	43.8→79.8	56.4→86.3
total	137.4→116.6	97.6→112.2	104.6→168.2	128.6→142.0	131.2→148.0	133.7→184.3

Experiments



The fine-tuning performance achieved by initializing from different offline algorithms



(a) From TD3+BC to SAC (b) From online DT to TD3 (c) From online DT to PPO

The fine-tuning performance achieved by transferring to three online algorithms from their heterogeneous offline algorithms.

Code