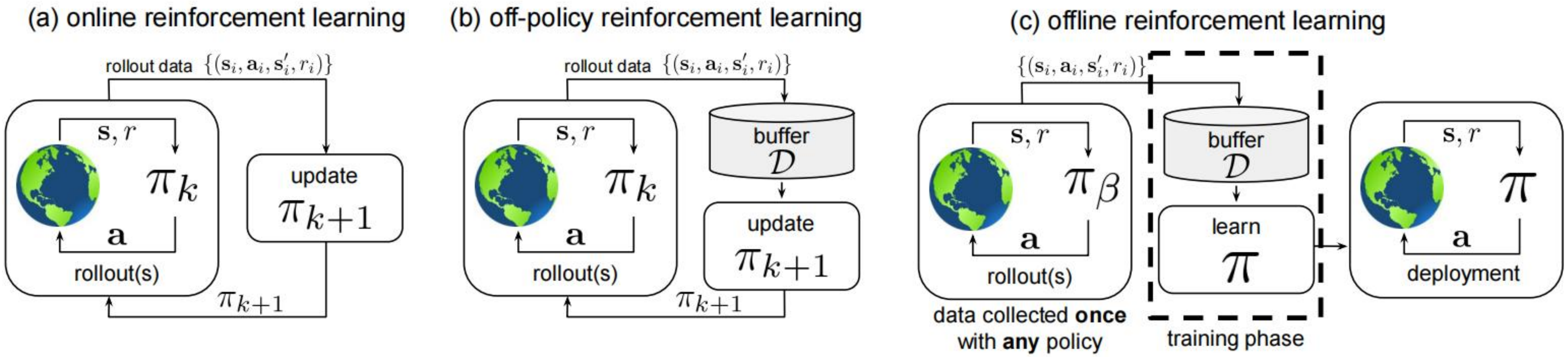
Offline RL Policies Should be Trained to be Adaptive

Dibya Ghosh [1] **Anurag Ajay** [2] **Pulkit Agrawal** [2] **Sergey Levine** [1]

ICML 2022

Background



Notions

MDP : $\mathcal{M} = (S, A, R, P, \rho, \gamma)$

Value function : $V(s)$

Action value function : $Q(s, a)$

Objective function of π : $J(\pi_\theta) = \mathbb{E}_{s_0 \sim \rho(s_0), a \sim \pi_\theta(\cdot|s_0)}[Q^{\pi_\theta}(s_0, a_0)] = \mathbb{E}_{\pi_\theta}\left[\sum_{t \geq 0} \gamma^t r(s_t, a_t)\right]$

Motivation

Uncertainty in offline RL represents agent's estimation of environment

Static datasets of offline RL $\Longrightarrow$ the learned policy is often conservative

The learned policy penalizes actions with high uncertainty to avoid OOD actions

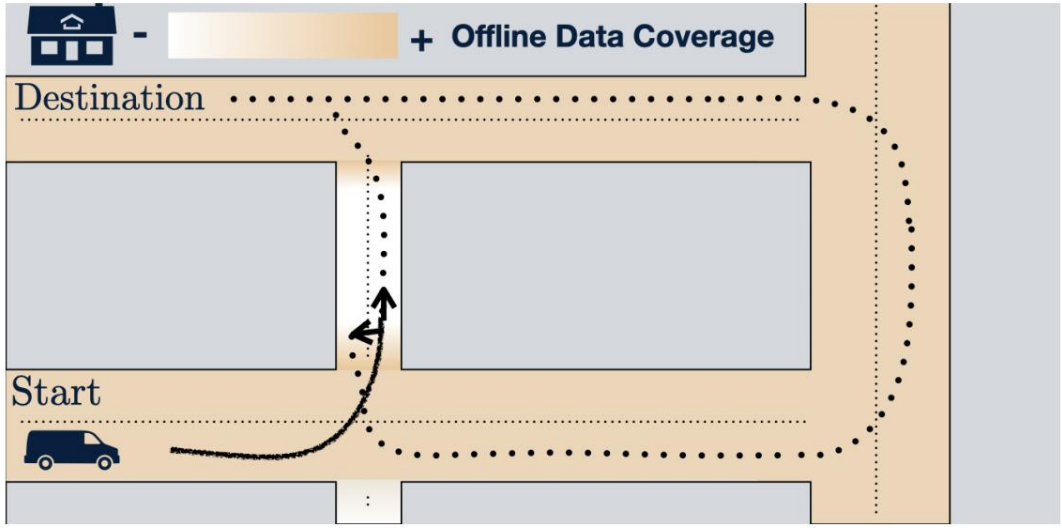
Use an ensemble of Q-networks

- $$J(\theta)=\mathbb{E}_{s,a\sim D}\left[\mathbb{E}_{Q^{\pi}\sim\mathcal{P}_D(\cdot)}[Q^{\pi}(s,a)]-\alpha U_{\mathcal{P}_D}(\mathcal{P}_D(\cdot))\right]$$
- take the minimum Q value while update Q-networks

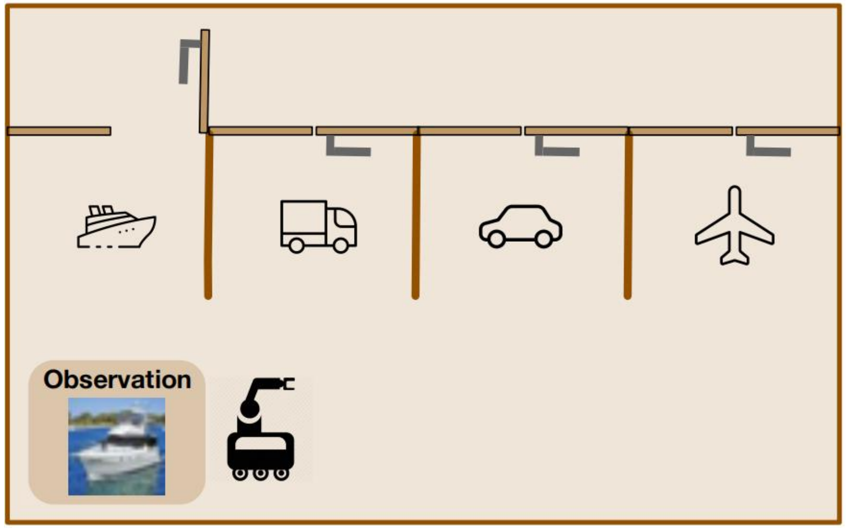
But it remains unclear whether conservative objectives are the best approach for designing offline RL algorithms.

Motivation

City navigation



Locked doors



Efficiency first : the side street
Conservative policy : the large road

Adaptive : try the side street and reverts to the large road if unknown circumstances arise

Standard offline RL : error prediction of image will try to open a locked door forever

Adaptive : try another door

Offline RL Policies Should be Trained to be Adaptive

Analysis

Insufficient data coverage → Many potential MDPs behave identically in the dataset, but differ on out-of-sample states and actions.

The dataset does not uniquely identify M^* of the true environment

So the dataset induces epistemic uncertainty about the identity of the MDP

From a Bayesian perspective:

Given a prior distribution over MDPs $P(\mathcal{M})$, then $P(\mathcal{M}|\mathcal{D}) \propto P(\mathcal{D})P(\mathcal{D}|\mathcal{M})$

Because the learned policy will be deployed into the true MDP, so the Bayesian objective is

$$J_{\text{Bayes}}(\pi) := \mathbb{E}_{\mathcal{M}\sim P(\mathcal{M}|\mathcal{D})}[J_{\mathcal{M}}(\pi)] \iff \text{the expected return of epistemic POMDP}$$

$$\text{where } J_{\mathcal{M}}(\pi) = \mathbb{E}_{\pi}\left[\sum\nolimits_{t\geq 0}\gamma^t r(s_t, a_t)\right]$$

Analysis

POMDP is defined by $\left(\bar{S}, A, \mathcal{O}, \bar{P}, O, r, \rho, \gamma\right)$

The epistemic POMDP : $J_{\mathcal{M}_{po}}(\pi) = J_{\text{Bayes}}(\pi)$; $\bar{s} := (s, \mathcal{M})$; $\bar{P}((s', \mathcal{M}') | (s, \mathcal{M}), a)$

$$O((s, \mathcal{M})) = s \; ; \; r((s, \mathcal{M}), a) = r_{\mathcal{M}}(s, a) \; ; \; \rho((s, \mathcal{M})) = P(\mathcal{M} | \mathcal{D}) \rho_{\mathcal{M}}(s)$$

Use the parlance of partial observability to describe how uncertainty induced by the offline dataset affects policy learning and evaluation process under a Bayesian viewpoint.

Proposition A.1 (Sub-optimality of Markovian policies and optimality of adaptiveness). *Let $n \in \mathbb{N}$. There are offline RL problem instances $(\mathcal{D}, p(\mathcal{M}))$ with n-state MDPs where the adaptive Bayes-optimal policy achieves $J_{Bayes}(\pi^*_{adaptive}) = -2n$ but the highest performing Markovian policy achieves return of a magnitude worse: $J_{Bayes}(\pi^*_{markov}) \leq -\frac{1}{2}n^2$.*

Method

The posterior distribution $P(\mathcal{M}|\mathcal{D})$ is unacquirable, so use $P(Q^{\pi}_{\mathcal{M}}|\mathcal{D})$ instead

(Because the value function $Q^{\pi}_{\mathcal{M}}$ entangles the necessary information about both dynamics and rewards for a given policy)

Difine relative MDP belief $\boldsymbol{b}(h)(\mathcal{M}) = \frac{P(\mathcal{M}|h,\mathcal{D})}{P(\mathcal{M}|\mathcal{D})}$

Traditional policy gradient $\nabla_\theta J_{\mathcal{M}}(\pi_\theta) = \mathbb{E}_{h\sim\pi}[\nabla_\theta \mathbb{E}_{a\sim\pi_\theta(\cdot|h)}[Q^{\pi}_{\mathcal{M}}(h,a)]]$

Bayesian policy gradient $\nabla_\theta J_{Bayes}(\pi_\theta) = \mathbb{E}_{h\sim\pi}[\nabla_\theta \mathbb{E}_{a\sim\pi_\theta(\cdot|s(h),b(h))}[\mathbb{E}_{M\sim P(\mathcal{M}|\mathcal{D})}[b(h)(\mathcal{M})Q^{\pi}_{\mathcal{M}}(h,a)]]$

The update of Q function :

$$Q^{\pi}_{\mathcal{M}}(s,\boldsymbol{b},a) = r(s,a) + \gamma \mathbb{E}_{\substack{s'\sim\mathcal{M} \\ a\sim\pi}}[Q^{\pi}_{\mathcal{M}}(s',\boldsymbol{b}',a)] \quad (5)$$

where $\boldsymbol{b}' := \mathrm{BeliefUpdate}(\boldsymbol{b},(s,a,r,s))$ *is the new relative MDP belief after witnessing* (s,a,r,s'),

$$\mathrm{BeliefUpdate}(\boldsymbol{b},(s,a,r,s))(\mathcal{M}) \propto p_{\mathcal{M}}(r,s'|s,a)\boldsymbol{b}(\mathcal{M}) \quad (6)$$

Method

$$\boldsymbol{b}(h)(\mathcal{M}) = \frac{P(\mathcal{M}|h,\mathcal{D})}{P(\mathcal{M}|\mathcal{D})} = \frac{P(h,\mathcal{D}|\mathcal{M})P(\mathcal{M})}{P(\mathcal{M}|\mathcal{D})P(h,\mathcal{D})} = \frac{P(h|\mathcal{D},\mathcal{M})P(\mathcal{D})P(\mathcal{M})}{P(\mathcal{M}|\mathcal{D})P(h|\mathcal{D})P(\mathcal{D})} = \frac{P(h|\mathcal{D},\mathcal{M})P(\mathcal{M})}{P(\mathcal{M}|\mathcal{D})P(h|\mathcal{D})}$$

Because $P(\mathcal{M}|\mathcal{D}), P(\mathcal{M})$ is fixed for different $\mathcal{M}$, and for different h, $P(h|\mathcal{D})$ is the same or similar

$$\boldsymbol{b}(h')(\mathcal{M}) \propto P(h'|\mathcal{D},\mathcal{M}) = P_{\mathcal{M}}(s',r|s,a)P(h|\mathcal{D},\mathcal{M})$$

$$\Rightarrow \boldsymbol{b}(h')(\mathcal{M}) \propto P_{\mathcal{M}}(s',r|s,a)\boldsymbol{b}(h)(\mathcal{M})$$

Without model, replace $P_{\mathcal{M}}(s',r|s,a)$ by $\log \hat{P}_{\mathcal{M}_k}(s',r|s,a) = -\left|\hat{Q}_k(s,\boldsymbol{b},a) - (r + \gamma \mathbb{E}_{a'\sim\pi}[\hat{Q}_k(s',\boldsymbol{b},a')])\right|$

Finally, the actor loss and critic loss are :

$$\mathcal{L}_{\text{critic}}(\hat{Q}_k) = \mathbb{E}_{(s,a,r,s')\sim\mathcal{D},\boldsymbol{b}\sim p(\boldsymbol{b})}\left[\left(\hat{Q}_k(s,\boldsymbol{b},a) - \left(r + \gamma \mathbb{E}_{a'\sim\pi(\cdot|s',\boldsymbol{b}')}\left[\hat{Q}_k(s',\boldsymbol{b}',a')\right]\right)\right)^2\right]$$

$$\mathcal{L}_{\text{actor}}(\pi) = -\mathbb{E}_{s\sim\mathcal{D},\boldsymbol{b}\sim p(\boldsymbol{b})}\left[\mathbb{E}_{a\sim\pi(\cdot|s,\boldsymbol{b})}\left[\sum_k \boldsymbol{b}_k \hat{Q}_k(s,\boldsymbol{b},a)\right]\right]$$

Method

Algorithm 1 Adaptive Policies with Ensembles of Value Functions (APE-V)

Receive input: dataset $\mathcal{D}$, number of ensemble members n
Initialize policy $\pi(\cdot|s, \boldsymbol{b}) : \mathcal{S} \times \Delta_n \to \Delta(\mathcal{A})$
Initialize ensemble of value functions $\{\hat{Q}_1, \dots \hat{Q}_n\}$, where $\hat{Q}_k(s, \boldsymbol{b}, a) : \mathcal{S} \times \Delta_n \times \mathcal{A} \to \mathbb{R}$ → $\boldsymbol{b}_0 = \left[\frac{1}{n} \cdots \frac{1}{n}\right]^{\mathrm{T}}$
while π has not converged **do**
Sample transition $(s, a, r, s') \sim \mathcal{D}$ from dataset and possible belief $\boldsymbol{b} \sim p(\boldsymbol{b})$
Approximate next-step belief $\boldsymbol{b}' = \text{BeliefUpdate}(\boldsymbol{b}, (s, a, r, s'))$ using Equation 9
Optimize value functions to minimize TD error taking into account the updated belief $\boldsymbol{b} \to \boldsymbol{b}'$

$$\min \mathcal{L}(\hat{Q}_k) := (\hat{Q}_k(s, \boldsymbol{b}, a) - (r + \gamma \mathbb{E}_{a' \sim \pi(\cdot|s', \boldsymbol{b}')}[\hat{Q}_k(s', \boldsymbol{b}', a')]|))^2 \quad \forall k \in \{1, \dots n\} \tag{7}$$

Optimize adaptive policy $\pi(\cdot|s, \boldsymbol{b})$ to maximize $\boldsymbol{b}$-weighted average of value functions

$$\max_{\pi(\cdot|s,\boldsymbol{b})} \mathbb{E}_{a_\pi \sim \pi}[\sum_k \boldsymbol{b}_k \hat{Q}_k(s, \boldsymbol{b}, a_\pi)] \tag{8}$$

end while

Algorithm 2 APE-V Test-Time Adaptation

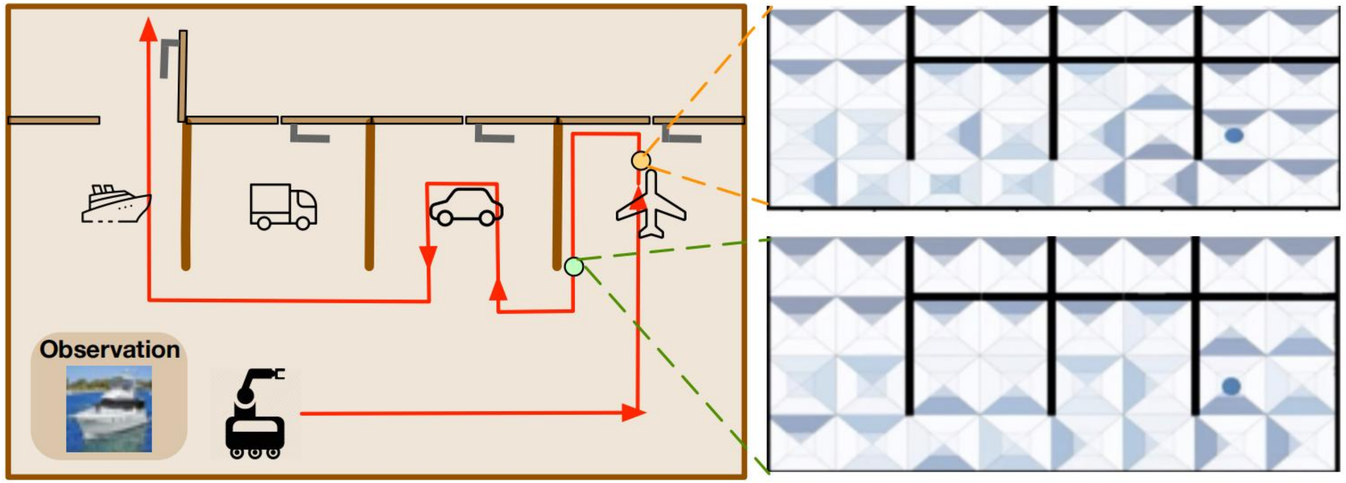
$s_0 = \text{ENV.RESET}()$
Initialize belief vector to uniform: $\boldsymbol{b}_0 = [\frac{1}{n}, \dots, \frac{1}{n}]^\top$
for environment step $t = 0, 1, \dots$ **do**
Sample action: $a_t \sim \pi(\cdot|s_t, \boldsymbol{b}_t)$
Act in environment: $r_t, s_{t+1} \leftarrow \text{ENV.STEP}(a_t)$
Update belief vector using new transition (Eq 9)
$\boldsymbol{b}_{t+1} = \text{BeliefUpdate}(\boldsymbol{b}_t, (s_t, a_t, r_t, s_{t+1}))$
end for

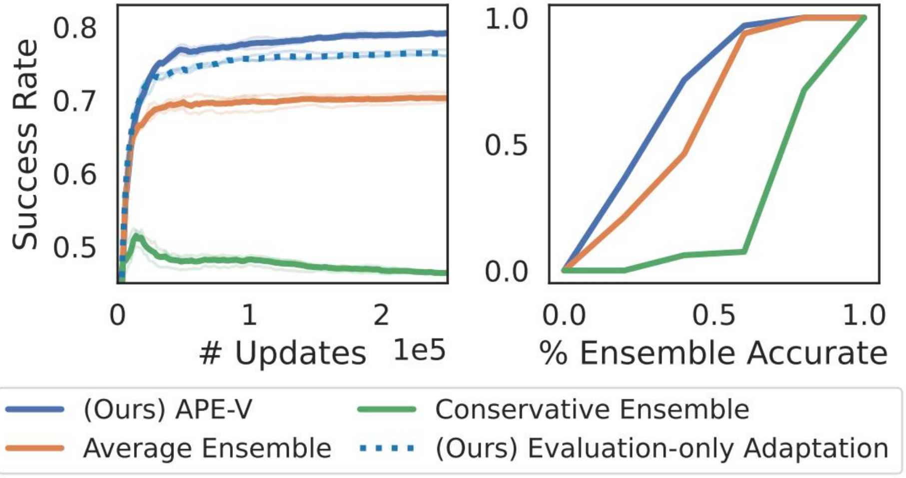
$$\text{BeliefUpdate}(\boldsymbol{b}, (s, a, r, s))(\mathcal{M}) \propto p_{\mathcal{M}}(r, s'|s, a)\boldsymbol{b}(\mathcal{M})$$

$$\log \hat{P}_{\mathcal{M}_k}(s', r|s, a) = -\left|\hat{Q}_k(s, \boldsymbol{b}, a) - (r + \gamma \mathbb{E}_{a' \sim \pi}[\hat{Q}_k(s', \boldsymbol{b}, a')])\right|$$

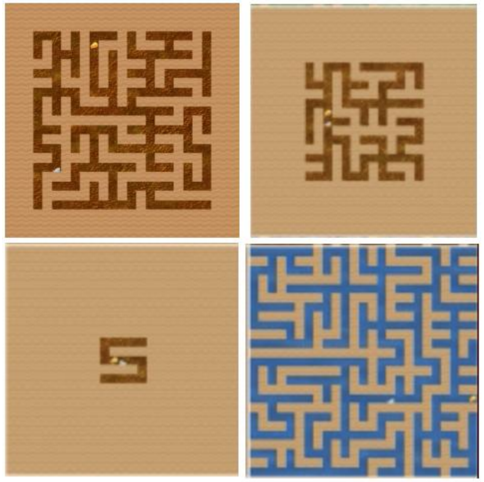
Experiment

Locked Doors with CIFAR10





Procgen Mazes



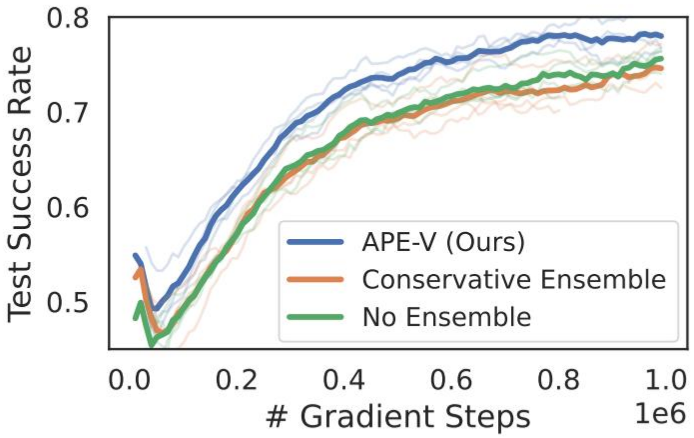


Table 2. Maze-solving success rates, averaged over 4 seeds.

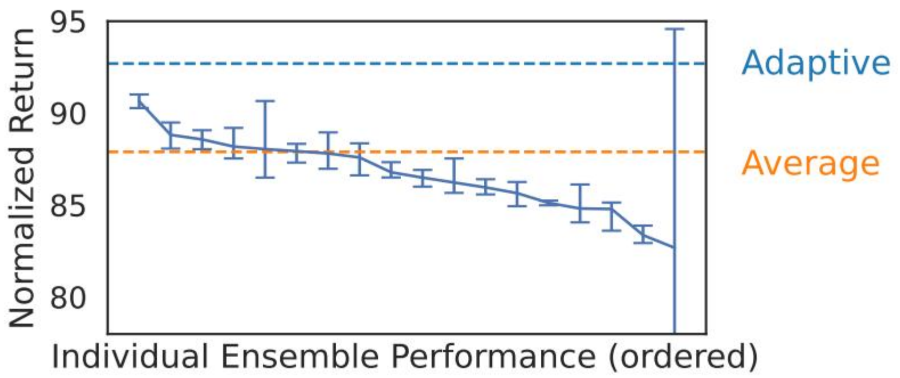
	200 Train Levels		**1000 Train Levels**	
	Train	Test	Train	Test
No Ensemble	0.96 ±0.02	0.24 ±.02	0.93 ±0.01	0.75 ±0.03
Conservative	0.96 ±0.02	0.23 ±0.02	0.93 ±0.01	0.75 ±0.04
APE-V	0.97 ±0.00	**0.31** ±0.04	0.92 ±0.01	**0.79** ±0.04

Experiment

D4RL benchmark

Task Name	BC	SAC (Haarnoja et al., 2018)	REM (Agarwal et al., 2020)	CQL (Kumar et al., 2020)	IQL (Kostrikov et al., 2021b)	SAC-N (An et al., 2021)	APE-V
halfcheetah-random	2.2±0.0	29.7±1.4	-0.8±1.1	**35.4**	31.3±3.5	29.8±1.6	29.9±1.1
halfcheetah-medium	43.2±0.6	55.2±27.8	-0.8±1.3	44.4	47.4±0.2	67.5±1.2	**69.1 ± 0.4**
halfcheetah-medium-expert	44.0±1.6	28.4±19.4	0.7±3.7	62.4	95.0±1.4	**102.7±1.5**	**101.4 ± 1.4**
halfcheetah-medium-replay	37.6±2.1	0.8±1.0	6.6±11.0	46.2	44.2±1.2	**63.9±0.8**	**64.6 ± 0.9**
hopper-random	3.7±0.6	9.9±1.5	3.4±2.2	10.8	5.3±0.6	**31.3±0.0**	**31.3±0.2x**
hopper-medium-expert	53.9±4.7	0.7±0.0	0.8±0.0	**111.0**	96.9±15.1	110.1±0.3	105.72 ± 3.7
hopper-medium-replay	16.6±4.8	7.4±0.5	27.5±15.2	48.6	94.7±8.6	**101.8±0.5**	98.5 ± 0.5
walker2d-random	1.3±0.1	0.9±0.8	6.9±8.3	7.0	5.4±1.7	**16.3±9.4**	**15.5±8.5**
walker2d-medium	70.9±11.0	-0.3±0.2	0.2±0.7	74.5	78.3±8.7	87.9±0.2	**90.3 ± 1.6**
walker2d-medium-expert	90.1±13.2	1.9±3.9	-0.1±0.0	98.7	109.1±0.2	**116.0±6.3**	110.0 ± 1.5
walker2d-medium-replay	20.3±9.8	-0.4±0.3	12.5±6.2	32.6	73.8±7.1	78.7±0.7	**82.9 ± 0.4**

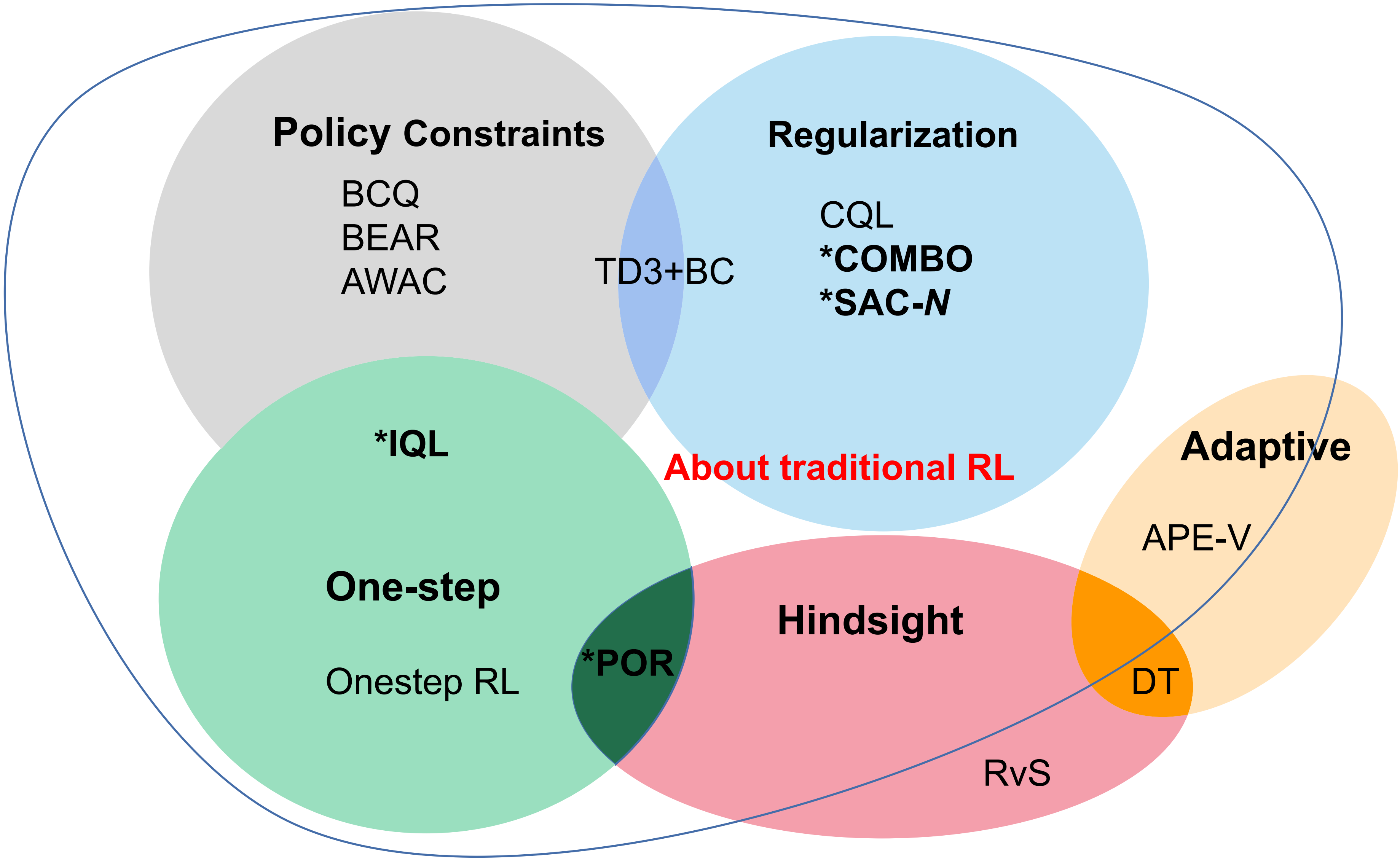
These tasks generally do not have data distributions that lead to multiple salient hypotheses



Adaptation within the episode indeed allows the policy to adapt to a better strategy than it may have started with.

Discussion

Sota algorithms in offline RL



Thanks