ACTIVE: Offline RL via Adaptive Imitation and In-sample V-Ensemble

Tianyuan Chen [1,2] Ronglong Cai [1] Faguo Wu [1,2*] Xiao Zhang [1,2*]

[1]Beihang University [2]Zhongguancun Laboratory

Introduction

- **Offline RL** learns policies from static datasets without environment interaction
- **Key Challenge:** Value estimation errors for out-of-distribution (OOD) actions
- **In-sample Learning:** Avoids OOD value queries by approximating in-sample maximum
- **Limitations:** Existing in-sample methods suffer from **over-regularization** on suboptimal datasets

Over-regularization on Mixed/Suboptimal Datasets

Value Learning: Gap between Theory and Practice

$$\mathcal{L}_V(\psi) = \mathbb{E}_{(s,a)\sim\mathcal{D}}[L_2^\tau(Q_{\hat{\theta}}(s,a) - V_\psi(s))], \quad \text{where} \quad L_2^\tau(u) = |\tau - \mathbb{1}(u<0)|u^2. \tag{1}$$

- Theory: $\tau \to 1$ better approximates **in-sample maximum**
- Practice: $\tau \to 1$ leads to **overestimation or divergence**

Policy Extraction: Conservativeness of AWR

$$\mathcal{L}_\pi(\phi) = \mathbb{E}_{(s,a)\sim\mathcal{D}}[\exp(\beta(Q_{\hat{\theta}}(s,a) - V_\psi(s)))\log \pi_\phi(a|s)]. \tag{2}$$

- AWR imposes KL-divergence constraint by minimizing $D_{\text{KL}}(\pi_{\text{imp}}||\pi_\phi)$
- AWR tends to **monotonically increase dataset likelihood** $\mathbb{E}_{(s,a)\sim\mathcal{D}}[\log \pi(a|s)]$

Overestimation with High Expectile τ

- In implicit TD backups, $V(s)$ can easily overfit the initial error of $Q(s, a)$
- Prevents using larger expectile τ or lower implicit regularization α
- Results in over-regularization on suboptimal datasets
- Conservative policy that closely mimics behavior policy

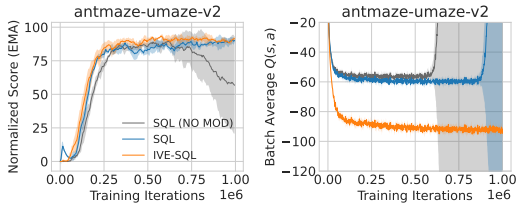


Figure: **Catastrophic errors in value estimation**

In-sample V-Ensemble (IVE)

Value Learning with Ensemble

$$\mathcal{L}_V^f(\psi_i) = \mathbb{E}_{(s,a)\sim\mathcal{D}}[f(Q_{\hat{\theta}}(s,a) - V_{\psi_i}(s))] \tag{3}$$

$$\mathcal{L}_Q(\theta) = \mathbb{E}_{(s,a,s')\sim\mathcal{D}}[(r(s,a) + \gamma \min_i V_{\psi_i}(s') - Q_\theta(s,a))^2] \tag{4}$$

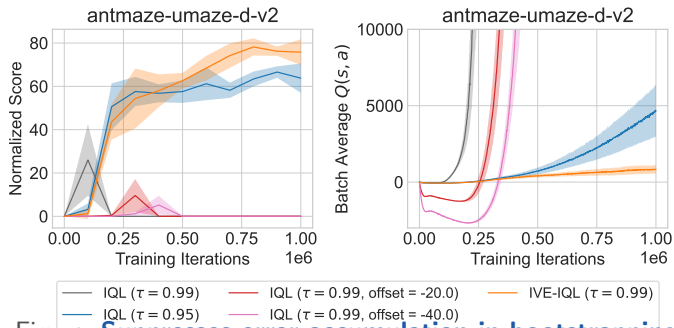


Figure: **Suppresses error accumulation in bootstrapping**

In-sample V-Ensemble (IVE)

Value Learning with Ensemble

$$\mathcal{L}_V^f(\psi_i) = \mathbb{E}_{(s,a)\sim\mathcal{D}}[f(Q_{\hat{\theta}}(s,a) - V_{\psi_i}(s))] \tag{5}$$

$$\mathcal{L}_Q(\theta) = \mathbb{E}_{(s,a,s')\sim\mathcal{D}}[(r(s,a) + \gamma \min_i V_{\psi_i}(s') - Q_\theta(s,a))^2] \tag{6}$$

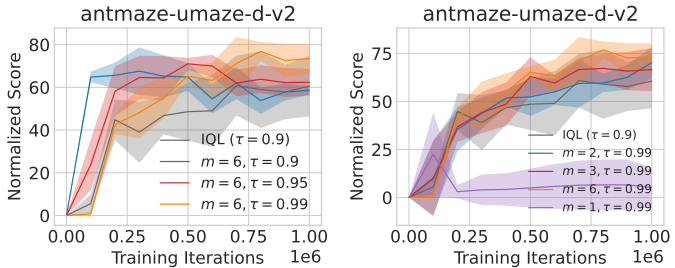


Figure: **Enables use of larger expectile τ (or lower regularization α)**

In-sample V-Ensemble (IVE)

Advantage Estimation

$$A(s,a) = Q_{\hat{\theta}}(s,a) - \mathbb{E}^c[V_{\psi_i}](s) \tag{7}$$

where $\mathbb{E}^c[V_{\psi_i}](s)$ is the c-th quantile of the ensemble

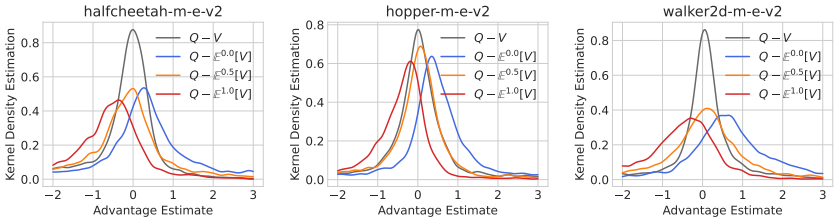


Figure: **Controls advantage sparsity** $s(Q, V; \mathcal{D}) = \mathbb{E}_{(s,a)\sim\mathcal{D}}[\mathbb{1}(Q(s,a) - V(s) \leq 0)]$ **with** c

In-sample V-Ensemble (IVE)

In-sample Epistemic Uncertainty

$$\mathbb{E}\left[\min_{1\leq i\leq m} V_i(s)\right] \approx \mu(s) - \Phi^{-1}\left(\frac{m-\frac{\pi}{8}}{m-\frac{\pi}{4}+1}\right)\sigma(s) \tag{8}$$

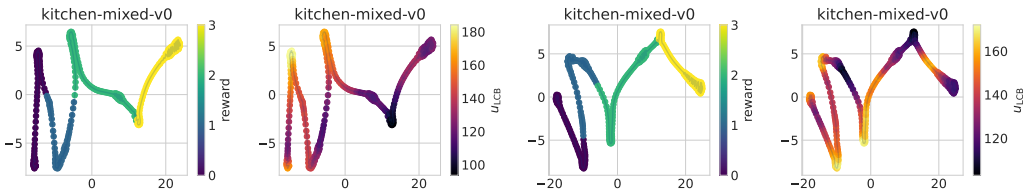


Figure: **Uncertainty estimation helps even when there are no OOD actions**

Theoretical Analysis (IVE)

- In the NTK regime, ensemble-based bootstrapping suppresses initial errors in value estimation
- For output dynamics of Q-function after t iterations:

$$Q^{(t+1)}(\mathcal{X}) = \left[1 + \cdots + \left(\gamma C_V C_{\text{imp}}\right)^t\right] R + O\left(\gamma^t \| C_V C_{\text{imp}} \|^t\right) \tag{9}$$

$$+ \gamma \left[1 + \cdots + \left(\gamma C_V C_{\text{imp}}\right)^{t-1}\right] (\Omega - B) \tag{10}$$

- $B = \sqrt{\mathbb{V}_{\text{ens}}[V_{\text{lin}}^{(t)}(\mathcal{O}')]}$ represents uncertainty penalty from the ensemble
- Larger ensemble size leads to larger B, suppressing the implicit optimization of π_{imp} that selects high $Q^{(0)}$ values

Adaptive Cloning Temperature (ACT)

Constrained Optimization Problem

$$\min_{\phi} -\mathbb{E}_{s\sim\mathcal{D},a\sim\pi_\phi}[Q_{\hat{\theta}}(s,a)] \tag{11}$$

$$\text{s.t. } \mathbb{E}_{(s,a)\sim\mathcal{D}}[\log \pi_\phi(a|s)] \geq \mathcal{H}_{\mathcal{D}} \tag{12}$$

Policy Extraction with Adaptive Cloning Temperature

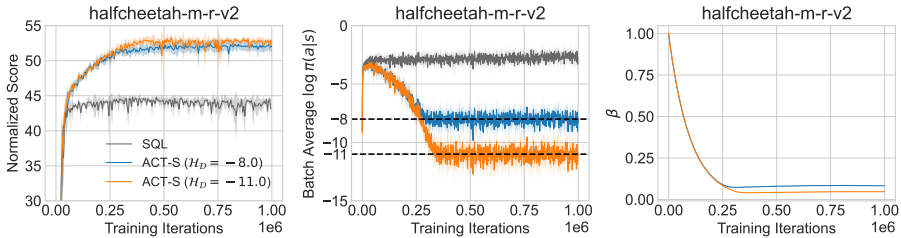
$$\mathcal{L}_\pi(\phi) = -\mathbb{E}_{s\sim\mathcal{D},a\sim\pi_\phi}[Q_{\hat{\theta}}(s,a)] - \beta\mathbb{E}_{(s,a)\sim\mathcal{D}}[w(s,a)\log \pi_\phi(a|s)] \tag{13}$$

$$\mathcal{L}(\beta) = \mathbb{E}_{(s,a)\sim\mathcal{D}}[\beta(\log \pi_\phi(a|s) - \mathcal{H}_{\mathcal{D}})] \tag{14}$$

where $w(s,a) = \exp(Q_{\hat{\theta}}(s,a) - \mathbb{E}^c[V_{\psi_i}](s))$.

Benefits of Adaptive Cloning Temperature

- Adaptively adjusts imitation strength using dual gradient descent
- Balances between regularization and generalization
- **Avoids monotonic increase** in dataset likelihood with AWR
- More discriminative on suboptimal data



The Complete Algorithm

Algorithm ACTIVE

Hyperparameters: $f = f_\alpha$, m, $\mathcal{H}_\mathcal{D}$, LR λ, λ_β, EMA η.
Initialize: ϕ, θ, $\hat{\theta}$, $\{\psi_i\}_{i=1}^m$, β, $\mathcal{D}$.
for each gradient step **do**

$$\psi_i \leftarrow \psi_i - \lambda \nabla_{\psi_i} \mathcal{L}_V^f(\psi_i)$$
$$\theta \leftarrow \theta - \lambda \nabla_\theta \mathcal{L}_Q(\theta)$$
$$\hat{\theta} \leftarrow (1-\eta)\hat{\theta} + \eta\theta$$
$$\beta \leftarrow \beta - \lambda_\beta \nabla_\beta \mathcal{L}(\beta)$$
$$\phi \leftarrow \phi - \lambda \nabla_\phi \mathcal{L}_\pi(\phi)$$

end for

Experimental Results

- ACTIVE outperforms baseline in-sample methods across datasets
- Particularly effective on **mixed/suboptimal** datasets

Dataset	CQL	TD3+BC	IQL	SQL	IQL (Reproduced)	SQL (Reproduced)	ACTIVE-I	ACTIVE-S
halfcheetah-m	44.0 ±0.8	48.3 ±0.3	47.4 ±0.2	48.3 ±0.2	47.4 ±0.1	47.8 ±0.0	50.6 ±0.1	**52.3 ±0.2**
hopper-m	58.5 ±2.1	59.3 ±4.2	66.2 ±5.7	75.5 ±3.4	64.4 ±2.7	63.8 ±2.3	80.1 ±3.8	**86.1 ±3.4**
walker2d-m	72.5 ±0.8	83.7 ±2.1	78.3 ±8.7	84.2 ±4.6	80.1 ±1.3	82.8 ±0.4	84.8 ±0.7	**87.2 ±1.0**
halfcheetah-m-r	45.5 ±0.5	44.6 ±0.5	44.2 ±1.2	44.8 ±0.7	43.6 ±0.2	43.9 ±0.3	51.1 ±0.3	**51.7 ±0.2**
hopper-m-r	95.0 ±6.4	60.9 ±18.8	94.7 ±8.6	101.7 ±3.3	86.0 ±13.4	85.6 ±12.3	101.1 ±1.6	**102.8±0.7**
walker2d-m-r	77.2 ±5.5	81.8 ±5.5	73.8 ±7.1	77.2 ±3.8	70.3 ±5.6	69.2 ±6.3	**85.4 ±3.5**	79.3 ±4.6
halfcheetah-m-e	90.7 ±4.3	90.7 ±4.3	86.7 ±5.3	**94.0 ±0.4**	88.2 ±1.6	91.2 ±1.2	93.2 ±0.4	92.9 ±1.0
hopper-m-e	105.4 ±6.8	98.0 ±9.4	91.5 ±14.3	**111.8 ±2.2**	98.7 ±7.9	104.9 ±5.4	89.2 ±10.9	109.9 ±2.0
walker2d-m-e	109.6 ±0.7	110.1 ±0.5	109.6 ±1.0	110.0 ±0.8	110.0 ±0.4	111.2 ±0.1	**112.0 ±0.4**	111.7 ±0.3
antmaze-u	74.0	78.6	87.5 ±2.6	**92.2 ±1.4**	85.8 ±3.2	90.0 ±3.1	91.6 ±1.4	90.8 ±1.5
antmaze-u-d	**84.0**	71.4	62.2 ±13.8	74.0 ±2.3	60.6 ±14.2	49.6 ±6.9	78.4 ±5.1	43.8 ±22.1
antmaze-m-p	61.2	10.6	71.2 ±7.3	**80.2 ±3.7**	76.2 ±6.2	72.8 ±3.2	74.6 ±5.5	78.8 ±1.7
antmaze-m-d	53.7	3.0	70.0 ±10.9	75.1 ±4.2	73.8 ±3.7	62.2 ±8.1	**75.2 ±2.2**	67.0 ±3.3
antmaze-l-p	15.8	0.2	39.6 ±5.8	50.2 ±4.8	48.2 ±6.5	38.2 ±5.2	51.2 ±5.8	**51.2 ±5.7**
antmaze-l-d	14.9	0.0	47.5 ±9.5	**52.3 ±5.2**	47.2 ±3.5	41.6 ±4.4	48.2 ±5.8	42.6 ±3.2
kitchen-c	43.8 ±11.2	-	61.4 ±9.5	**76.4 ±8.7**	67.0 ±3.4	60.4 ±1.3	68.2 ±3.6	66.5 ±3.3
kitchen-p	49.8 ±10.1	-	46.1 ±8.5	72.5 ±7.4	58.8 ±8.7	70.8 ±4.1	69.6 ±3.0	**73.5 ±3.0**
kitchen-m	51.0 ±6.5	-	52.8 ±4.5	67.4 ±5.4	47.9 ±3.7	46.9 ±11.8	52.1 ±2.9	**73.2 ±1.4**

Experimental Results

Ablation Studies

- The impact of $\mathbb{E}^c[V_i]$ on policy performance varies across datasets

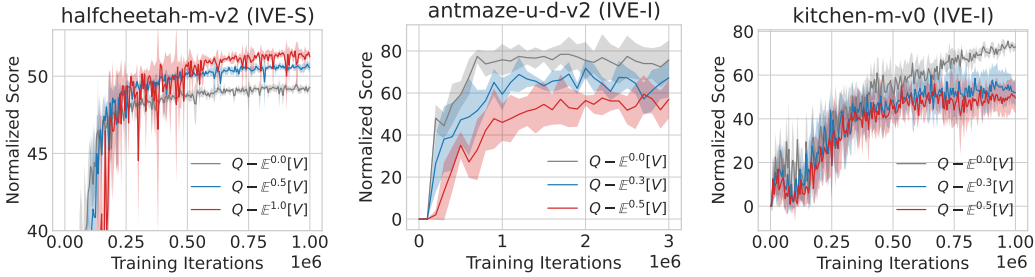


Figure: Sensitivity to quantile c of V-ensemble

Experimental Results

Ablation Studies

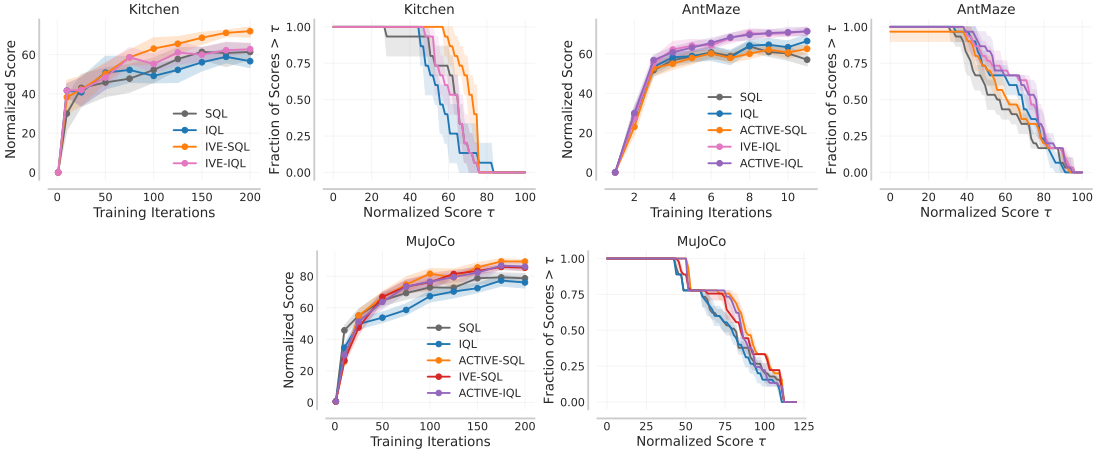


Figure: Ablation studies on Kitchen, Antmaze, and Mujoco datasets

Conclusion and Takeaways

- ACTIVE effectively mitigates over-regularization in in-sample learning methods
- V-ensemble prevents value function overfitting
- Adaptive cloning temperature balances between regularization and generalization
- Strong empirical performance across various offline RL datasets
- Future work: Incorporating out-of-sample learning for better generalization

Paper URL

Thank You!