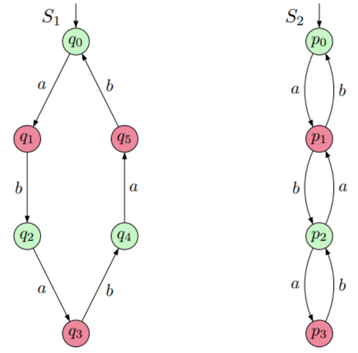
Understanding and Addressing the Pitfalls of Bisimulation-based Representations in Offline Reinforcement Learning

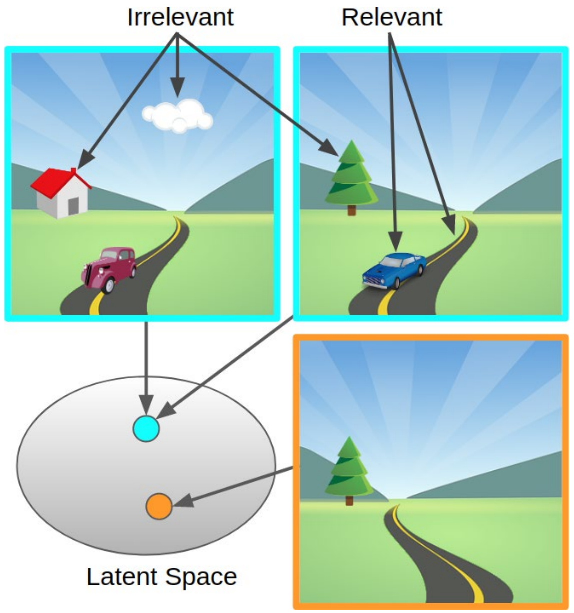
Hongyu Zang[1], Xin Li[1], Leiji Zhang[1],
Yang Liu[2], Baigui Sun[2], Riashat Islam[3],
Remi Tachet des Combes[4], Romain Laroche

[1]Beijing Institute of Technology, China
[2]Alibaba Group, China
[3]Mila, McGill University, Canada
[4]Wayve, UK

Bisimulation



Bisimilar states and bisimilar labeled transition systems



(Image source: Zhang et al. 2021)

Theorem 1 (Castro 2019):

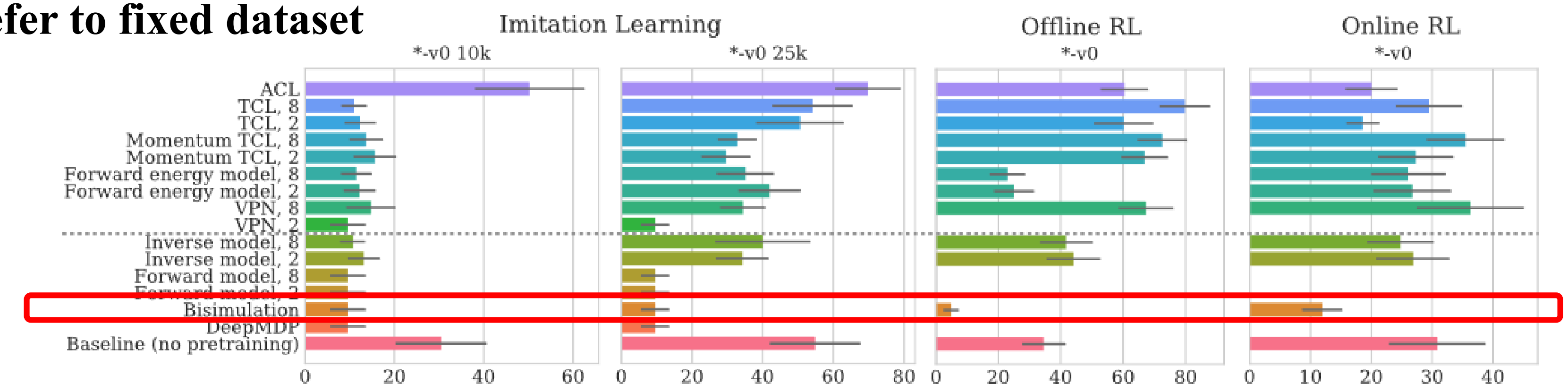
Define $\mathcal{F}^\pi : \mathcal{M} \to \mathcal{M}$ by $\mathcal{F}^\pi(d)(s,t) = | \mathcal{R}_s^\pi - \mathcal{R}_t^\pi | + \gamma \mathcal{W}_1(d) \left(\mathcal{P}_s^\pi, \mathcal{P}_t^\pi\right)$, then $\mathcal{F}^\pi$ has a least fixed point $d_\sim^\pi$, and $d_\sim^\pi$ is a π-bisimulation metric.

- DBC [Zhang et al. 2021]
- MICo [Castro et al. 2021]
- SimSR [Zang et al. 2022]
- …

Perform pretty well in Online settings!

Bisimulation in Offline RL

When refer to fixed dataset



(Image source: Yang et al. 2021)

Motivations

- While bisimulation-based approaches hold promise for learning robust state representations for Reinforcement Learning (RL) tasks, their efficacy in offline RL tasks has not been up to par.
- Recent studies suggest that bisimulation-based algorithms yield significantly poorer results on Offline tasks compared to a variety of (self-)supervised objectives.

Contributions

- We investigate the pitfalls of directly applying the bisimulation principle in Offline settings.
- We propose theoretically motivated modifications, including an expectile-based operator and a tailored reward scaling strategy.
- We demonstrate superior performance through empirical studies on D4RL and Visual D4RL

Formal Usage of Bisimulation

Goal: Approximate the fixed point of bisimulation measurement

bisimulation error Δ_ϕ^π: $\Delta_\phi^\pi(s_i, s_j) := |G_\phi^\pi(s_i, s_j) - G_\sim^\pi(s_i, s_j)|$

minimizing the distance between the approximation G_ϕ^π and the fixed point $G_\sim^\pi$

× **Obstacle**: the fixed point $G_\sim^\pi$ is unobtainable.

Lemma (Lifted MDP): The bisimulation-based update operator for an MDP is the Bellman evaluation operator for a specific lifted MDP.

✓ **Solution**:

Define ***bisimulation Bellman residual*** ϵ_ϕ^π as:

$$\epsilon_\phi^\pi(s_i, s_j) := |G_\phi^\pi(s_i, s_j) - \mathcal{F}^\pi G_\phi^\pi(s_i, s_j)|,$$

and given the connection between the bisimulation operator and MDP,
we can minimize bisimulation Bellman residual instead.

Formal Usage of Bisimulation

bisimulation error Δ_ϕ^π: $\Delta_\phi^\pi(s_i,s_j) := |G_\phi^\pi(s_i,s_j) - G_\sim^\pi(s_i,s_j)|$

bisimulation Bellman residual ϵ_ϕ^π: $\epsilon_\phi^\pi(s_i,s_j) := |G_\phi^\pi(s_i,s_j) - \mathcal{F}^\pi G_\phi^\pi(s_i,s_j)|$,

Theorem 3. ***(Bisimulation error upper-bound).*** *Let $\mu_\pi(s)$ denote the stationary distribution over states, let $\mu_\pi(\cdot,\cdot)$ denote the joint distribution over synchronized pairs of states (s_i, s_j) sampled independently from $\mu_\pi(\cdot)$. For any state pair $(s_i, s_j) \in \mathcal{S} \times \mathcal{S}$, the bisimulation error $\Delta_\phi^\pi(s_i, s_j)$ can be upper-bounded by a sum of expected bisimulation Bellman residuals ϵ_ϕ^π:*

$$\Delta_\phi^\pi(s_i, s_j) \le \frac{1}{1-\gamma}\mathbb{E}_{(s_i',s_j')\sim\mu_\pi}\left[\epsilon_\phi^\pi(s_i', s_j')\right]. \tag{5}$$

Proposition 4. ***(The expected bisimulation residual is not sufficient over incomplete datasets).*** *If there exists states s_i' and s_j' not contained in dataset $\mathcal{D}$, where the occupancy $\mu_\pi(s_i'|s_i, a_i) > 0$ and $\mu_\pi(s_j'|s_j, a_j) > 0$ for some $(s_i, s_j) \sim \mu_\pi$, then there exists a bisimulation measurement G_ϕ^π and a constant $C > 0$ such that*

- *For all $(\hat{s}_i, \hat{s}_j) \in \mathcal{D}$, the bisimulation Bellman residual $\epsilon_\phi^\pi(\hat{s}_i, \hat{s}_j) = 0$.*
- *There exists $(s_i, s_j) \in \mathcal{D}$, such that the bisimulation error $\Delta_\phi^\pi(s_i, s_j) = C$.*

Modifications of Bisimulation in Offline RL

Two improvements:

- Expectile-based Bisimulation Operator

Online version, sample-based

$$\mathcal{F}^{\pi} G^{\pi}(s_i, s_j) = |r_{s_i}^{\pi} - r_{s_j}^{\pi}| + \gamma \mathbb{E}_{\substack{s_i' \sim T_{s_i}^{\pi} \\ s_j' \sim T_{s_j}^{\pi}}} [G^{\pi}(s_i', s_j')]$$

Offline version, sample-based

$$(\mathcal{F}_{\tau}^{\pi_\beta} G_{\phi}^{\pi_\beta})(s_i, s_j) := \underset{G_{\phi}^{\pi_\beta}}{\operatorname{argmin}} \ \mathbb{E}_{\substack{a_i \sim \pi_\beta(\cdot \mid s_i) \\ a_j \sim \pi_\beta(\cdot \mid s_j)}} [\tau [\hat{\epsilon}]_+^2 + (1-\tau)[-\hat{\epsilon}]_+^2],$$

$$\hat{\epsilon} = \mathbb{E}_{\substack{s_i' \sim T_{s_i}^{\pi_\beta} \\ s_j' \sim T_{s_j}^{\pi_\beta}}} [\underbrace{|r(s_i, a_i) - r(s_j, a_j)| + \gamma G_{\bar{\phi}}^{\pi_\beta}(s_i', s_j')}_{\text{target } G} - G_{\phi}^{\pi_\beta}(s_i, s_j)]\}$$

τ is used to balance a trade-off between behavior and optimal

Modifications of Bisimulation in Offline RL

Two improvements:

- Reward Scaling

Given a more general form of the bisimulation operator:

$$\mathcal{F}^{\pi} G(s_i, s_j) = c_{\tau} \cdot |r^{\pi}_{s_i} - r^{\pi}_{s_j}| + c_k \cdot \mathbb{E}^{\pi}_{s'_i, s'_j}[G(s'_i, s'_j)]$$

We can derive

$$\begin{aligned} G^{\pi}_{\sim}(s_i, s_j) &= \mathcal{F}^{\pi} G^{\pi}_{\sim}(s_i, s_j) = c_r \cdot |r^{\pi}_{s_i} - r^{\pi}_{s_j}| + c_k \cdot \mathbb{E}^{\pi}_{s_i', s_j'}[G^{\pi}_{\sim}(s_i', s_j')] \\ &\leq c_r \cdot (R_{\max} - R_{\min}) + c_k \cdot \mathbb{E}^{\pi}_{s_i', s_j'}[G^{\pi}_{\sim}(s_i', s_j')] \\ &\leq c_r \cdot (R_{\max} - R_{\min}) + c_k \cdot \max_{s_i', s_j'} G^{\pi}_{\sim}(s_i', s_j'). \end{aligned}$$

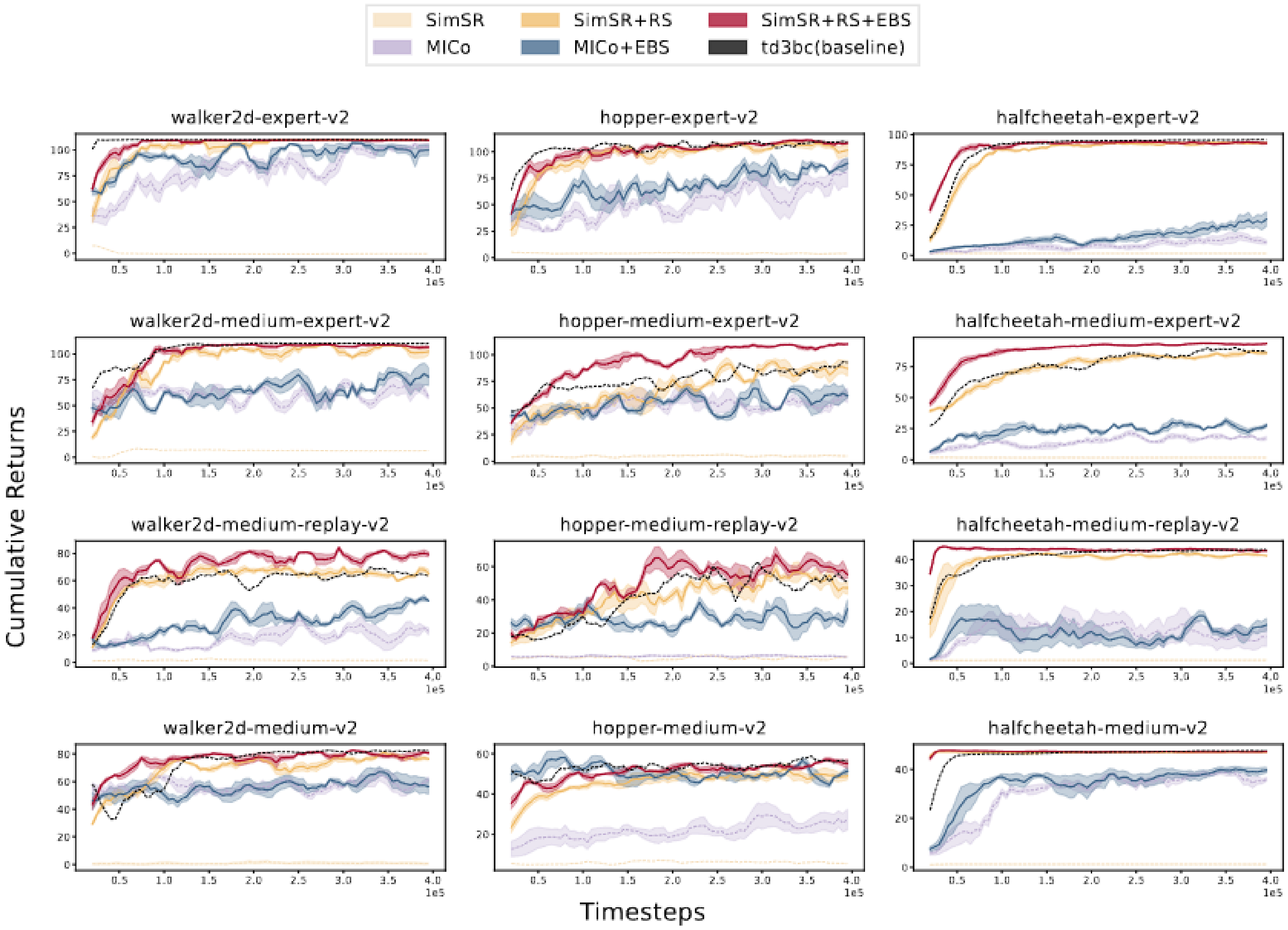
And

Theorem 8. *(Value bound based on on-policy bisimulation measurements in terms of approximation error). Given an MDP $\widetilde{\mathcal{M}}$ constructed by aggregating states in an ω-neighborhood, and an encoder ϕ that maps from states in the original MDP $\mathcal{M}$ to these clusters, the value functions for the two MDPs are bounded as*

$$\left|V^{\pi}(s) - \widetilde{V}^{\pi}(\phi(s))\right| \leq \frac{2\omega + \hat{\Delta}}{c_r(1-\gamma)}. \tag{11}$$

where $\hat{\Delta} := \|\hat{G}^{\pi}_{\sim} - \hat{G}^{\pi}_{\phi}\|_{\infty}$ is the approximation error.

Experiments



Experiments

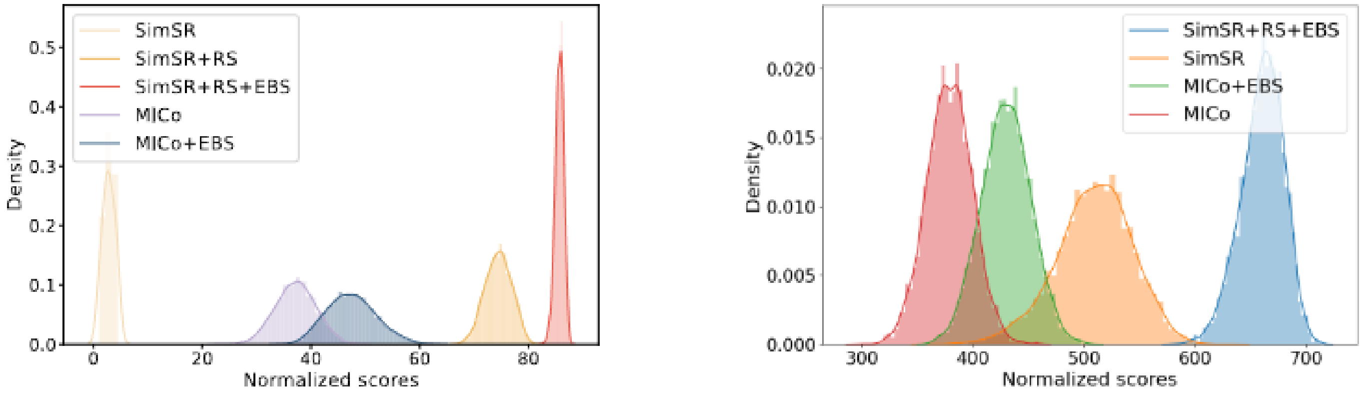


Figure 3: Bootstrapping distributions for uncertainty in IQM (*i.e.* inter-quartile mean) measurement on D4RL tasks (left) and visual D4RL tasks (right), following from the performance criterion in [2].

Table 1: Performance comparison with several other baselines on V-D4RL benchmark, averaged on 3 random seeds.

Dataset	CURL	DRIMLC	HOMER	ICM	MICo → MICo+EBS	SimSR → SimSR+RS+EBS
cheetah-run-medium	392	**524**	475	365	177 → 449 ($\nearrow$ 272)	391 → 491($\nearrow$ 100)
walker-walk-medium	452	425	439	358	450 → 447 (—)	443 → **480**($\nearrow$ 37)
cheetah-run-medium-replay	271	395	306	251	335 →357 ($\nearrow$ 22)	374 → **462**($\nearrow$ 88)
walker-walk-medium-replay	265	235	**283**	167	207 → 240 ($\nearrow$ 33)	197 → 240($\nearrow$ 43)
cheetah-run-medium-expert	348	403	383	280	282 → 341 ($\nearrow$ 59)	360 → **547**($\nearrow$ 187)
walker-walk-medium-expert	729	399	781	606	586 → 635($\nearrow$ 49)	755 → **845**($\nearrow$ 90)
cheetah-run-expert	200	310	218	237	308 → 331($\nearrow$ 23)	409 → **454**($\nearrow$ 45)
walker-walk-expert	769	427	686	**850**	370 → 447 ($\nearrow$ 77)	578 → 580 (—)
total	3426	3118	3571	3114	2715 → 3253 ($\nearrow$ 538)	3507 → **4043** ($\nearrow$ 536)

Check our paper for …

- Detailed description of our proposed method
- Theoretical guarantees
- More empirical results

paper

code