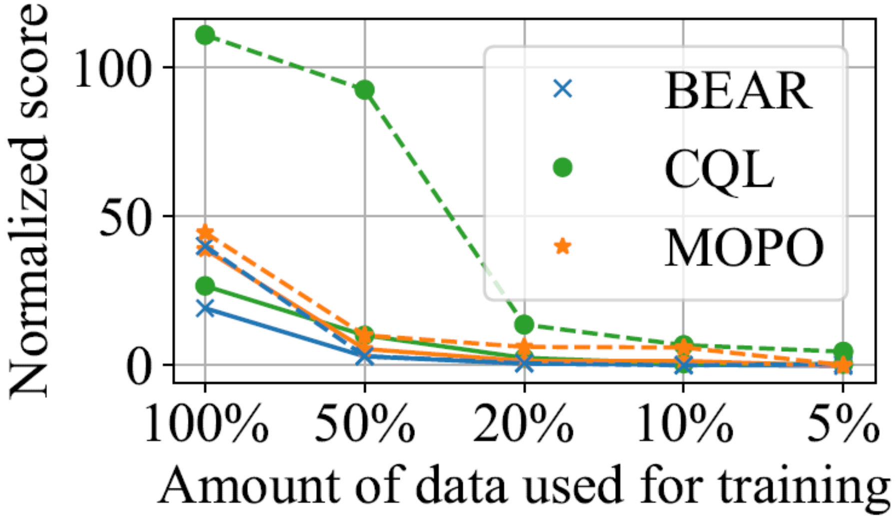
DARA: Dynamics-Aware Reward Augmentation in Offline Reinforcement Learning

Presenter: Jinxin Liu (liujinxin@westlake.edu.cn)
Co-Authors: Hongyin Zhang, Donglin Wang

Offline Reinforcement Learning (RL)

Offline reinforcement learning:
learning from the previously collected dataset.



Offline-data-hungry!

Collecting a large offline dataset for one specific task over one specific environment is costly and laborious.

Offline Domain (Dynamics) Adaptation

- Source → Target

Offline Domain (Dynamics) Adaptation

Limited target offline data: $\mathcal{D}$

$$\{(\mathbf{s}, \mathbf{a}, r, \mathbf{s}') \sim d_{\mathcal{D}}(\mathbf{s})\pi_b(\mathbf{a}|\mathbf{s})r(\mathbf{s}, \mathbf{a})T(\mathbf{s}'|\mathbf{s}, \mathbf{a})\}$$

Source offline data: $\mathcal{D}'$

$$\{(\mathbf{s}, \mathbf{a}, r, \mathbf{s}') \sim d_{\mathcal{D}'}(\mathbf{s})\pi_{b'}(\mathbf{a}|\mathbf{s})r(\mathbf{s}, \mathbf{a})T'(\mathbf{s}'|\mathbf{s}, \mathbf{a})\}$$

Assumption:

- Same state space
- Same action space
- Same reward function
- Different transition dynamics

- <u>Deterministic transition dynamics</u>

Definition 2 *(Dynamics shift) Let* $\hat{M} := (\mathcal{S}, \mathcal{A}, r, \hat{T}, \rho_0, \gamma)$ *be the empirical MDP estimated from* $\mathcal{D}$. *To evaluate a policy* π *for* $M := (\mathcal{S}, \mathcal{A}, r, T, \rho_0, \gamma)$ *with offline dataset* $\mathcal{D}$, *we say that the dynamics shift (between* $\mathcal{D}$ *and* M*) in offline RL happens if there exists at least one transition pair* $(\mathbf{s}, \mathbf{a}, \mathbf{s}') \in \{(\mathbf{s}, \mathbf{a}, \mathbf{s}') : d^{\pi}_{\hat{M}}(\mathbf{s})\pi(\mathbf{a}|\mathbf{s})\hat{T}(\mathbf{s}'|\mathbf{s}, \mathbf{a}) > 0\}$ *such that* $\hat{T}(\mathbf{s}'|\mathbf{s}, \mathbf{a}) \neq T(\mathbf{s}'|\mathbf{s}, \mathbf{a})$.

Lemma 2 *Dynamics shift produces that* $\mathcal{B}^{\pi}_{\mathcal{D}'}Q(\mathbf{s}, \mathbf{a}) \neq \mathcal{B}^{\pi}_{M}Q(\mathbf{s}, \mathbf{a})$ *for some* $(\mathbf{s}, \mathbf{a})$ *in* S'_{π}.

Dynamics-Aware Reward Augmentation

Lemma 2 *Dynamics shift produces that* $\mathcal{B}^{\pi}_{\mathcal{D}'}Q(\mathbf{s},\mathbf{a}) \neq \mathcal{B}^{\pi}_{M}Q(\mathbf{s},\mathbf{a})$ *for some* $(\mathbf{s},\mathbf{a})$ *in* S'_{π}.

Resort an additional compensation $\Delta_{\hat{T}',T}$ such that

$$\mathcal{B}^{\pi}_{\mathcal{D}'}Q(\mathbf{s},\mathbf{a}) + \Delta_{\hat{T}',T}(\mathbf{s},\mathbf{a}) = \mathcal{B}^{\pi}_{M}Q(\mathbf{s},\mathbf{a})$$

Algorithm 1 Framework for Dynamics-Aware Reward Augmentation (DARA)

Require: Target offline data $\mathcal{D}$ (reduced) and source offline data $\mathcal{D}'$

1: Learn classifiers (q_{sas} and q_{sa}) that distinguish source data $\mathcal{D}'$ from target data $\mathcal{D}$. (See Appendix A.1.3)

2: Set dynamics-aware $\Delta r(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1}) = \log \frac{q_{\text{sas}}(\text{source}|\mathbf{s}_t,\mathbf{a}_t,\mathbf{s}_{t+1})}{q_{\text{sas}}(\text{target}|\mathbf{s}_t,\mathbf{a}_t,\mathbf{s}_{t+1})} - \log \frac{q_{\text{sa}}(\text{source}|\mathbf{s}_t,\mathbf{a}_t)}{q_{\text{sa}}(\text{target}|\mathbf{s}_t,\mathbf{a}_t)}$.

3: Modify rewards for all $(\mathbf{s}_t, \mathbf{a}_t, r_t, \mathbf{s}_{t+1})$ in $\mathcal{D}'$: $r_t \leftarrow r_t - \eta\Delta r$.

4: Learn policy with $\{\mathcal{D} \cup \mathcal{D}'\}$ using prior model-free or model-based offline RL algorithms.

Dynamics-Aware Reward Augmentation

Lemma 2 *Dynamics shift produces that* $\mathcal{B}^{\pi}_{\mathcal{D}'}Q(\mathbf{s},\mathbf{a}) \neq \mathcal{B}^{\pi}_{M}Q(\mathbf{s},\mathbf{a})$ *for some* $(\mathbf{s},\mathbf{a})$ *in* S'_{π}.

Resort an additional compensation $\Delta_{\hat{T}',T}$ such that

$$\mathcal{B}^{\pi}_{\mathcal{D}'}Q(\mathbf{s},\mathbf{a}) + \Delta_{\hat{T}',T}(\mathbf{s},\mathbf{a}) = \mathcal{B}^{\pi}_{M}Q(\mathbf{s},\mathbf{a})$$

Algorithm 1 Framework for Dynamics-Aware Reward Augmentation (DARA)

Require: Target offline data $\mathcal{D}$ (reduced) and source offline data $\mathcal{D}'$

1: Learn classifiers (q_{sas} and q_{sa}) that distinguish source data $\mathcal{D}'$ from target data $\mathcal{D}$. (See Appendix A.1.3)

2: Set dynamics-aware $\Delta r(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1}) = \log \frac{q_{\text{sas}}(\text{source}|\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1})}{q_{\text{sas}}(\text{target}|\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}_{t+1})} - \log \frac{q_{\text{sa}}(\text{source}|\mathbf{s}_t, \mathbf{a}_t)}{q_{\text{sa}}(\text{target}|\mathbf{s}_t, \mathbf{a}_t)}$.

3: Modify rewards for all $(\mathbf{s}_t, \mathbf{a}_t, r_t, \mathbf{s}_{t+1})$ in $\mathcal{D}'$: $r_t \leftarrow r_t - \eta\Delta r$.

4: Learn policy with $\{\mathcal{D} \cup \mathcal{D}'\}$ using prior model-free or model-based offline RL algorithms.

$\Delta_{\hat{T}',T}$ discourages the learning from these offline transitions that are likely in source but are unlikely in target.

Experiments

	Body Mass Shift	10T	1T	1T+10S w/o Aug.	1T+10S DARA	10T	1T	1T+10S w/o Aug.	1T+10S DARA	10T	1T	1T+10S w/o Aug.	1T+10S DARA
		BEAR				BRAC-p				AWR			
Hopper	Random	11.4	1.0 ↓	4.6 ↑	8.4 ↑	11.0	10.9 ↓	9.6 ↓	**11.0** ↑	10.2	10.3 ↑	3.4 ↓	4.5 ↑
Hopper	Medium	52.1	0.8 ↓	0.9 ↑	1.6 ↑	32.7	29.0 ↓	29.2 ↑	**32.9** ↑	35.9	30.9 ↓	20.8 ↓	28.9 ↑
Hopper	Medium-R	33.7	1.3 ↓	18.2 ↑	**34.1** ↑	0.6	5.4 ↑	20.1 ↑	**30.8** ↑	28.4	8.8 ↓	4.1 ↓	4.2 ↑
Hopper	Medium-E	96.3	0.8 ↓	0.6 ↓	1.2 ↑	1.9	34.5 ↑	32.3 ↓	**34.7** ↑	27.1	27.0 ↓	26.8 ↓	**26.6** ↓
		BCQ				CQL				MOPO			
Hopper	Random	10.6	10.6 ↓	8.3 ↓	**9.7** ↑	10.8	10.6 ↓	10.2 ↓	**10.4** ↑	11.7	4.8 ↓	2.0 ↓	2.1 ↑
Hopper	Medium	54.5	37.1 ↓	25.7 ↓	38.4 ↑	58.0	43.0 ↓	44.9 ↑	**59.3** ↑	28.0	4.1 ↓	5.0 ↑	10.7 ↑
Hopper	Medium-R	33.1	9.3 ↓	28.7 ↑	**32.8** ↑	48.6	9.6 ↓	1.4 ↓	3.7 ↑	67.5	1.0 ↓	5.5 ↑	8.4 ↑
Hopper	Medium-E	110.9	58 ↓	75.4 ↑	84.2 ↑	98.7	59.7 ↓	53.6 ↓	**99.7** ↑	23.7	1.6 ↓	4.8 ↑	5.8 ↑
		BEAR				BRAC-p				AWR			
Walker2d	Random	7.3	1.5 ↓	3.1 ↑	3.2 ↑	-0.2	0.0 ↑	1.3 ↑	**3.2** ↑	1.5	1.3 ↓	2.0 ↑	**2.4** ↑
Walker2d	Medium	59.1	-0.5 ↓	0.6 ↑	0.3 ↓	77.5	6.4 ↓	70.0 ↑	**78.0** ↑	17.4	14.8 ↓	17.1 ↑	**17.2** ↑
Walker2d	Medium-R	19.2	0.7 ↓	6.5 ↑	7.3 ↑	-0.3	8.5 ↑	9.9 ↑	**18.6** ↑	15.5	7.4 ↓	1.6 ↓	1.5 ↓
Walker2d	Medium-E	40.1	-0.1 ↓	1.5 ↑	2.3 ↑	76.9	20.6 ↓	64.1 ↑	**77.5** ↑	53.8	35.5 ↓	52.5 ↑	**53.3** ↑
		BCQ				CQL				MOPO			
Walker2d	Random	4.9	1.8 ↓	4.5 ↑	**4.8** ↑	7.0	1.7 ↓	3.2 ↑	3.4 ↑	13.6	-0.2 ↓	-0.1 ↑	-0.1 ↓
Walker2d	Medium	53.1	32.8 ↓	50.9 ↑	**52.3** ↑	79.2	42.9 ↓	80.0 ↑	**81.7** ↑	17.8	7.0 ↓	5.7 ↓	11.0 ↑
Walker2d	Medium-R	15.0	6.9 ↓	14.9 ↑	**15.1** ↑	26.7	4.6 ↓	0.8 ↓	2.0 ↑	39.0	5.1 ↓	3.1 ↓	14.2 ↑
Walker2d	Medium-E	57.5	32.5 ↓	55.2 ↑	**57.2** ↑	111.0	49.5 ↓	63.5 ↑	93.3 ↑	44.6	5.3 ↓	5.5 ↑	17.2 ↑

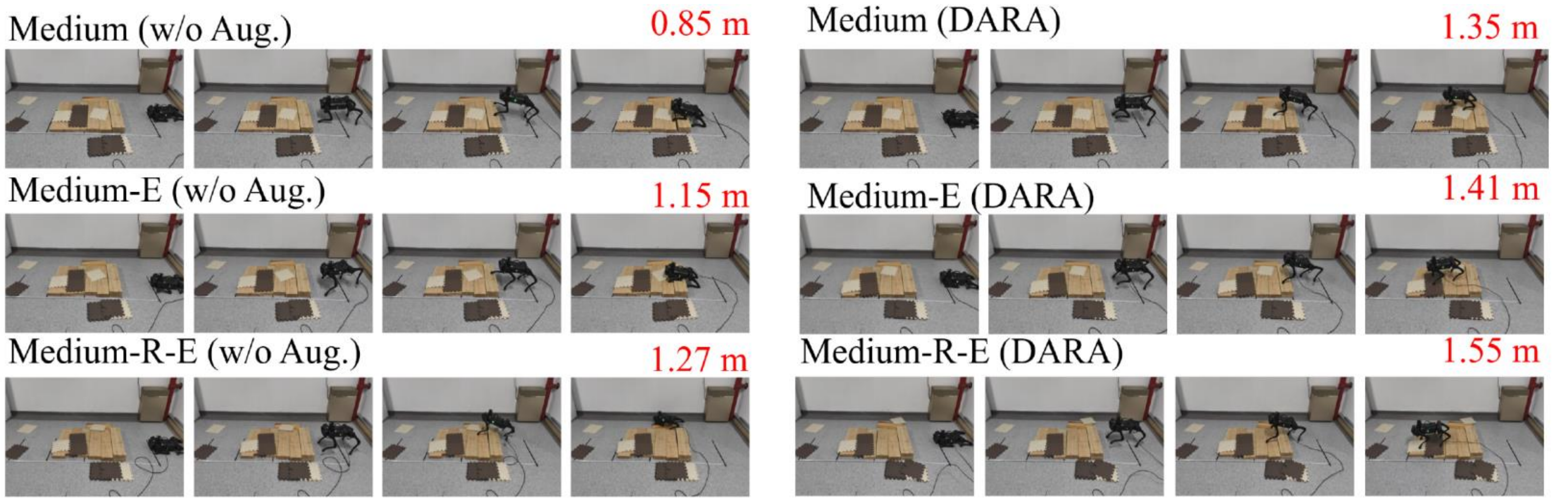
DARA can enable an adaptive policy with reduced offline data in target.

Experiments

	Body Mass Shift	Tune	DARA	Tune	DARA	Tune	DARA	Tune	DARA	Tune	DARA	$\pi_p\hat{T}$	$\hat{T}\pi_p$
Hopper		BEAR		BRAC-p		BCQ		CQL		MOPO		MABE	
	Random	0.8	8.4 ↑	6.0	11.0 ↑	8.8	9.7 ↑	**31.6**	10.4 ↓	0.7	2.1 ↑	10.6	9.0
	Medium	0.8	1.6 ↑	22.7	32.9 ↑	31.7	38.4 ↑	44.5	**59.3** ↑	0.7	10.7 ↑	48.8	23.1
	Medium-R	0.7	**34.1** ↑	14.7	30.8 ↑	27.5	32.8 ↑	1.3	3.7 ↑	0.6	8.4 ↑	17.1	20.4
	Medium-E	0.9	1.2 ↑	19.2	34.7 ↑	85.9	84.2 ↓	47.6	**99.7** ↑	2.2	5.8 ↑	28.1	38.9
Walker2d		BEAR		BRAC-p		BCQ		CQL		MOPO		MABE	
	Random	**6.6**	3.2 ↓	3.9	3.2 ↓	4.7	4.8 ↑	1.1	3.4 ↑	0.1	-0.1 ↓	6.0	-0.2
	Medium	0.3	0.3 ↓	76.0	78.0 ↑	28.4	52.3 ↑	72.3	**81.7** ↑	-0.2	11.0 ↑	30.1	56.7
	Medium-R	1.2	7.3 ↑	10.0	**18.6** ↑	10.4	15.1 ↑	1.8	2.0 ↑	0.0	14.2 ↑	13.3	12.5
	Medium-E	2.4	2.3 ↓	74.5	77.5 ↑	22.7	57.2 ↑	68.6	**93.3** ↑	7.3	17.2 ↑	43.7	82.7

Comparison with cross-domain baselines.

(BCQ)	w/o Aug.	DARA
Medium	0.85	1.35 ↑
Medium-E	1.15	1.41 ↑
Medium–R-E	1.27	1.55 ↑



Sim2real: Deployment on the obstructive and dynamic environment of BCQ.

Conclusion

1. The characterization of the dynamics shift in offline RL and the derivation of dynamics-aware reward augmentation (DARA) framework built on prior model-free and model-based formulations.

2. With only modest amounts of target offline data, we show that DARA-based offline methods can acquire an adaptive policy for the target tasks and achieve better performance compared to baselines in both simulated and real-world tasks.