A Unified Framework for Alternating Offline Model Training and Policy Learning

Shentao Yang[1], Shujian Zhang[1], Yihao Feng[2], Mingyuan Zhou[1]

[1]The University of Texas at Austin, [2]Salesforce Research

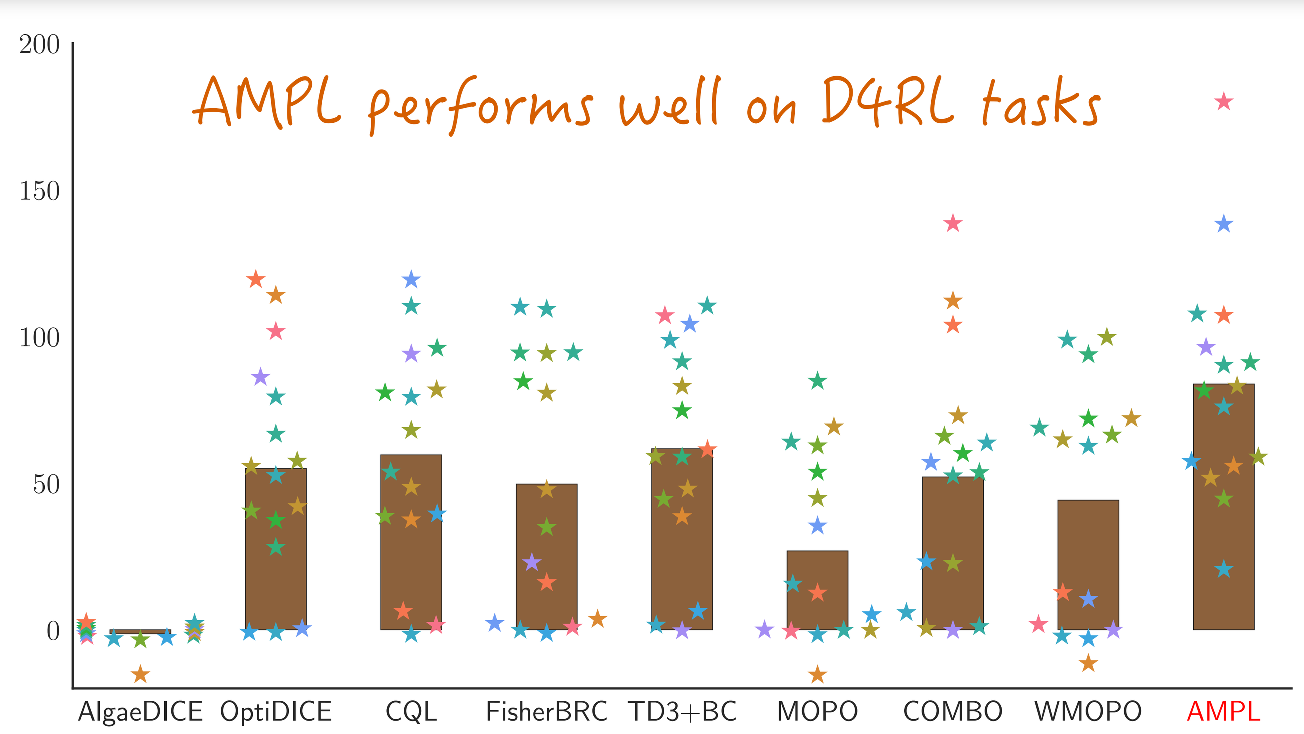
Proposed Method Sketch

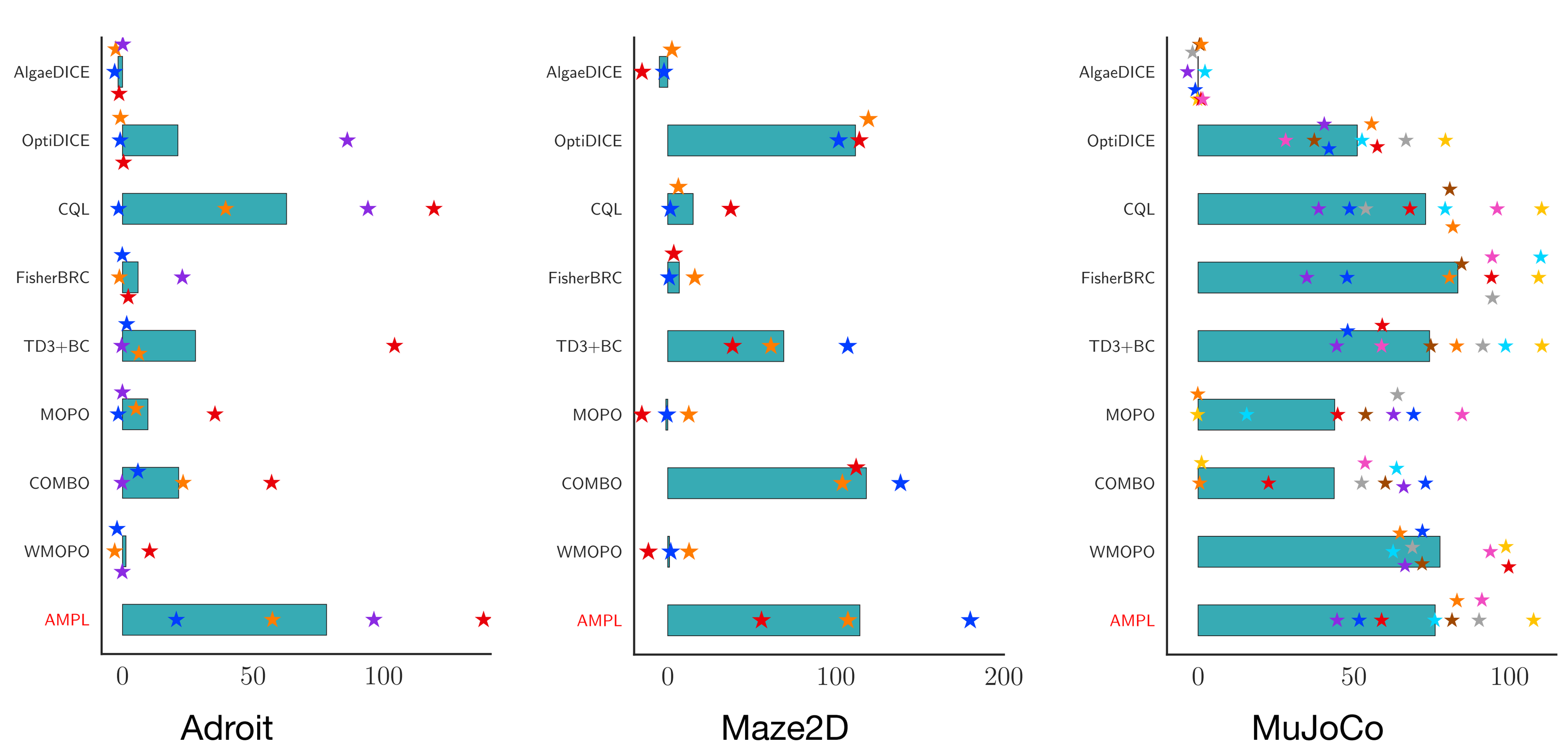
- **Motivation**: model training = MLE ≠ improve policy = model usage.

$$\min_{\pi, \widehat{P}} C \cdot \sqrt{D_\pi(P^*, \widehat{P})} \geq \left| J(\pi, P^*) - J(\pi, \widehat{P}) \right|$$

- Jointly train and to minimize an upper bound of the evaluation error.
- Fixed , $\approx$ only on state-actions visited by .
- Fixed , optimize with a regularization based on .

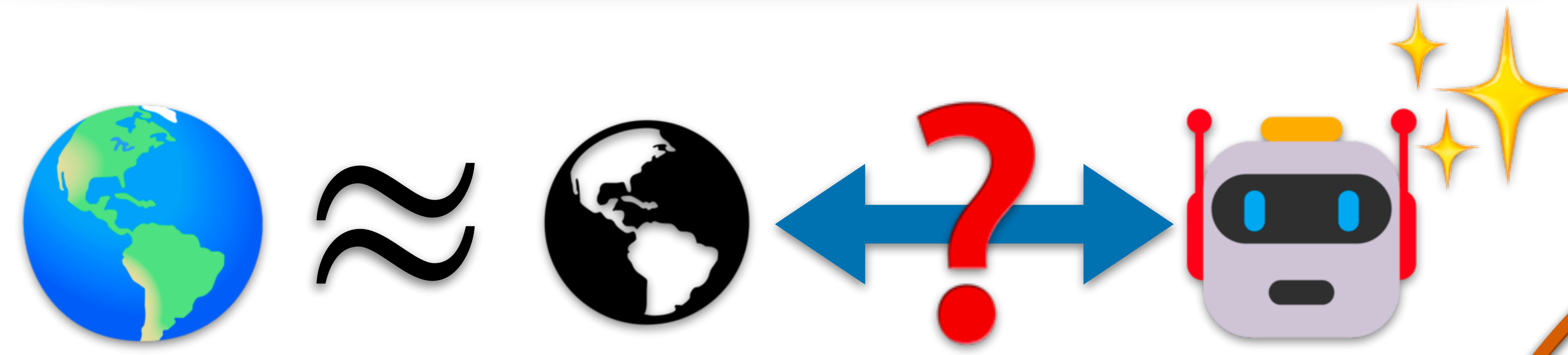
Results: AMPL





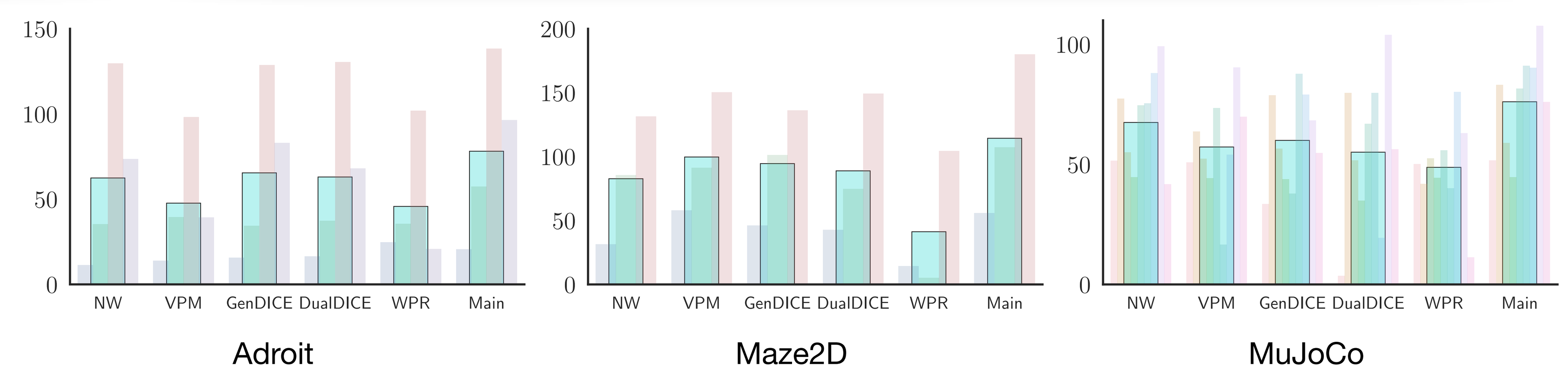
- Learn well on the MuJoCo datasets.
- Robust and good results on the challenging Adroit and Maze2D datasets.

Background

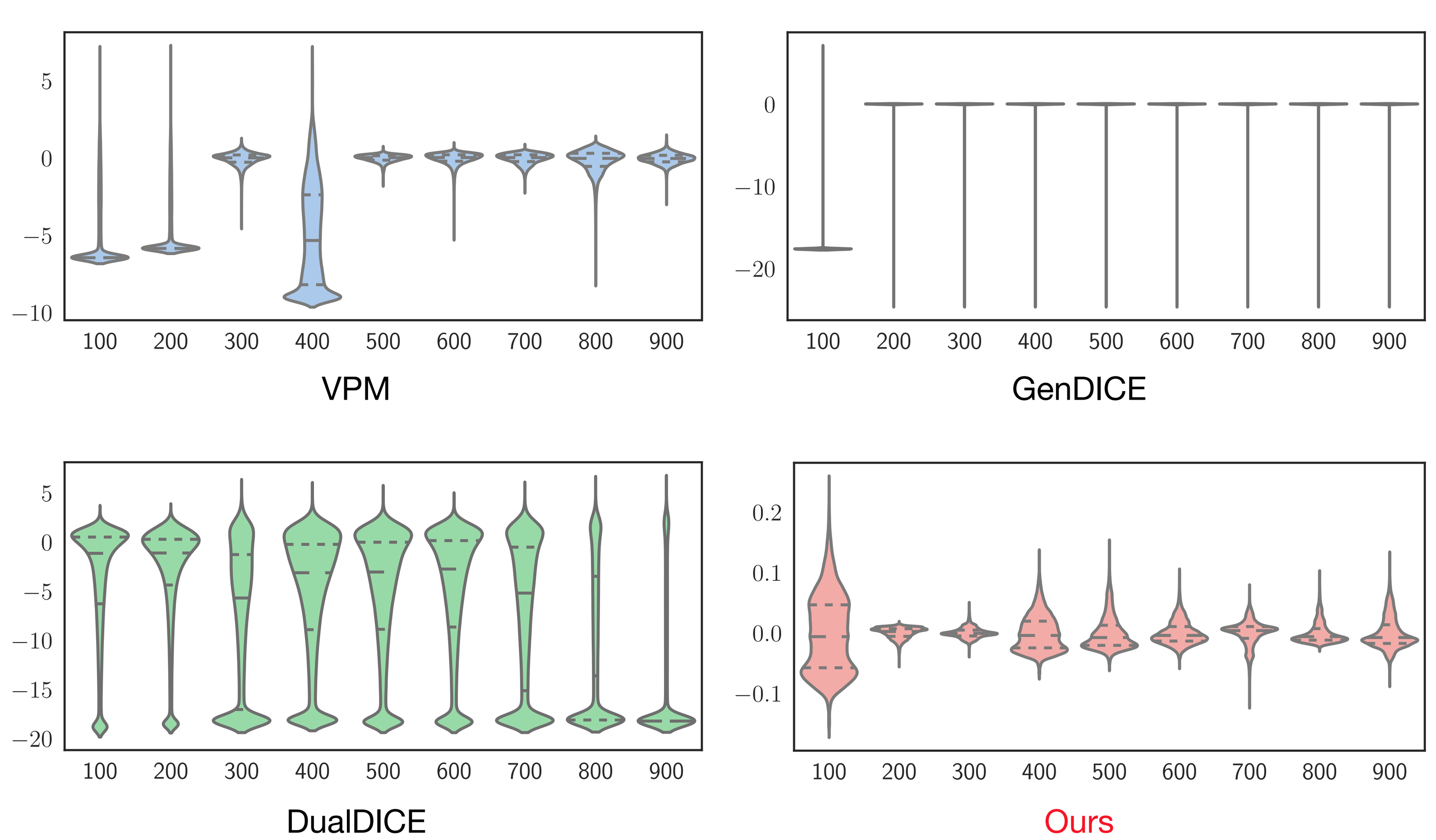


- Most offline MBRL: pre-train a fixed dynamic model on .
  - Objective: MLE — "simply a mimic of the world."
  - Usage: improve the policy.
- **Objective mismatch**: model training ≠ model usage.
  - Especially when is limited and is hard to learn.

Results: Ablation Study



- No Weights (NW): training only at the beginning using MLE.
- VPM, GenDICE, and DualDICE: other $\omega(s,a)$ estimation methods.
  - Can be unstable to provide good $\omega(s,a)$ for training.



Alternating Model-Policy Learning (AMPL)

- A tractable upper bound for the evaluation error

$$\left| J(\pi, P^*) - J(\pi, \widehat{P}) \right| \leq C \cdot \sqrt{D_\pi(P^*, \widehat{P})}, \quad \text{with}$$

$$D_\pi(P^*, \widehat{P}) \triangleq \mathbb{E}_{(s,a)\sim d^{P^*}_{\pi_b,\gamma}} \left[\omega(s,a) \, \mathrm{KL}\left(P^*(s' \mid s,a)\, \pi_b(a' \mid s') \,||\, \widehat{P}(s' \mid s,a)\, \pi(a' \mid s') \right) \right],$$

- where π_b is the behavior policy , $d^{P^*}_{\pi_b,\gamma}$ is the offline-data distribution ,

$\omega(s,a) \triangleq \dfrac{d^{P^*}_{\pi,\gamma}(s,a)}{d^{P^*}_{\pi_b,\gamma}(s,a)}$ is the density ratio between and visitation freq. of .

- Fix , train the **model** by $\ell(\widehat{P}) = -\mathbb{E}_{(s,a,s')\sim d^{P^*}_{\pi_b,\gamma}} \left[\omega(s,a) \log \left\{ \widehat{P}(s' \mid s,a) \right\} \right]$
- Given $\omega(s,a)$, a stable weighted MLE.

- Lower-bound of **policy** performance: $J\left(\pi, \widehat{P}\right) - C \cdot \sqrt{D_\pi(P^*, \widehat{P})}$.
- Fix , empirically helpful to construct the regularizer by:
  - Removing $\sqrt{\cdot}$, applying a further relaxation before changing KL ➡ JSD

$$D_\pi(P^*, \widehat{P}) \leq C'' \cdot \mathrm{KL}\left(P^*(s' \mid s,a)\, \pi_b(a' \mid s')\, d^{P^*}_{\pi_b,\gamma}(s,a) \,||\, \widehat{P}(s' \mid s,a)\, \pi(a' \mid s')\, d^{P^*}_{\pi_b,\gamma}(s)\, \pi(a \mid s) \right).$$

- **Density-ratio** training: Fixed-point-style simple MSE, ~~saddle-point optimization~~.

$$\mathbb{E}_{(s,a)\sim d^{P^*}_{\pi_b,\gamma}} \left[\omega(s,a) \cdot Q^{\widehat{P}}_\pi(s,a) \right] = \gamma\, \mathbb{E}_{\substack{(s,a,s') \sim d^{P^*}_{\pi_b,\gamma} \\ a' \sim \pi(\cdot \mid s')}} \left[\omega(s,a) \cdot Q^{\widehat{P}}_\pi(s',a') \right] + (1-\gamma)\, \mathbb{E}_{\substack{s \sim \mu_0(\cdot) \\ a \sim \pi(\cdot \mid s)}} \left[Q^{\widehat{P}}_\pi(s,a) \right].$$

- Based on primal-dual relation between $\omega(s,a)$ and Q-function in OPE.

Summary

- **Goal**: close the mismatched model objectives in offline MBRL.
- **Method**: offline Alternating Model-Policy Learning.

QR code for the full paper!

QR code for the GitHub Repo!