Offline Imitation from Observation via Primal Wasserstein State Occupancy Matching

Kai Yan, Alexander G. Schwing, Yu-Xiong Wang

ICML International Conference On Machine Learning

ILLINOIS

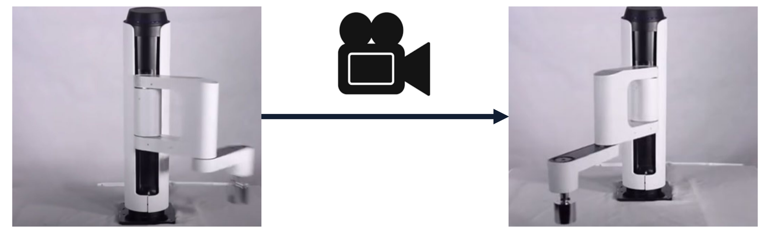
https://t.ly/yKi9V

Motivation

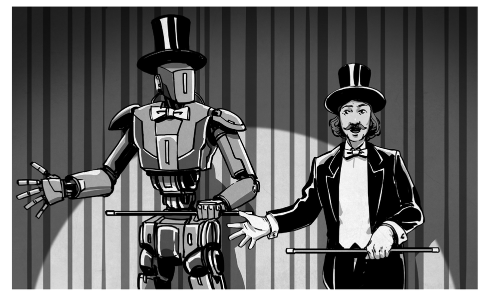
Goal: Offline imitation Learning from Observation (LfO)

Why from observations?

- Expert data are expensive, and action could be missing



Learning from video



Embodiment difference

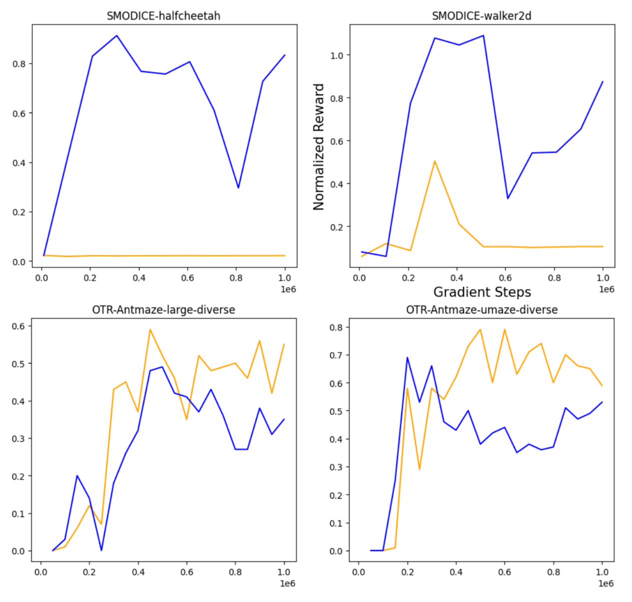
- Learning with few expert states + non-expert, mixed-quality state-action data with unknown optimality

Why Wasserstein distance?

- Prior works mostly use f-divergences, ignoring geometric distributions
- e.g., minimize state(-pair) occupancy KL [1, 2]

Why Primal Wasserstein?

- Most Wasserstein-based methods use **Rubinstein dual**, which limits underlying distance metric to be Euclidean
- Underlying metric is crucial to Wasserstein-based offline imitation; selecting a good metric is important
- PWIL [3] uses only a surrogate of primal distance



OTR [4]: cosine distance (orange) vs. Euclidean (blue)

Key Papers

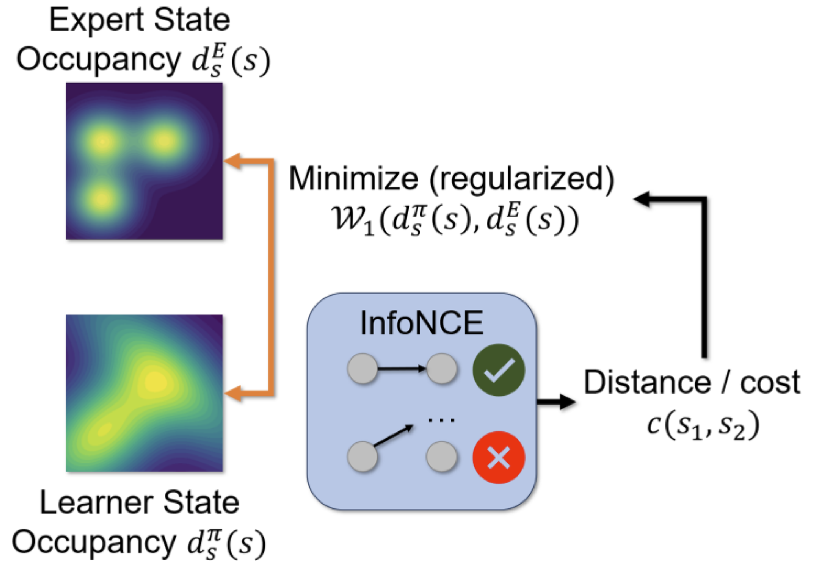
[1] Y. J. Ma et al. Smodice: Versatile offline imitation learning via state occupancy matching. In ICML, 2022.
[2] G. hyeong Kim et al. Lobsdice: Offline learning from observation via stationary distribution correction estimation. In NeurIPS, 2022.
[3] R. Dadashi et al. Primal wasserstein imitation learning. In ICLR, 2021.
[4] Y. Luo et al. Optimal Transport for offline imitation learning. In ICLR, 2023.

Formulation

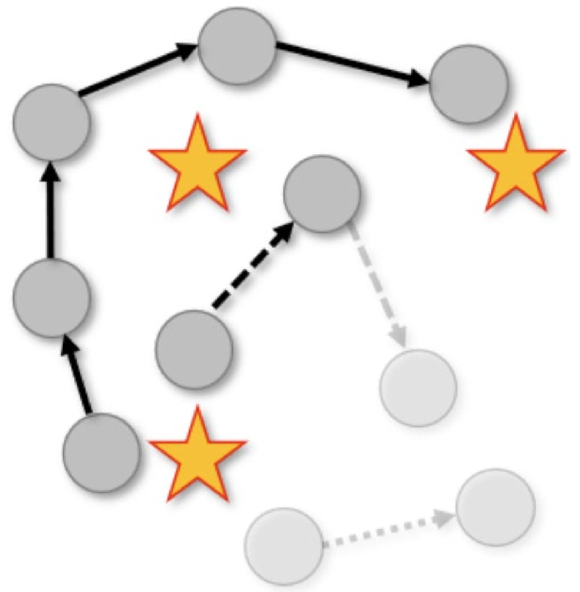
Primal problem:

$$\min_{\pi} \mathcal{W}(d_s^{\pi}, d_s^{E}) + \text{KL regularizers} \quad \text{s.t.} \quad \pi \text{ is feasible}$$

- $\mathcal{W}$ = 1-Wasserstein distance, d_s^E is the expert's state occupancy, d_s^{π} is state occupancy of learner's policy π



Wasserstein Optimization



Weighted Behavior Cloning

Contrastive learning for distance metric: weighted sum of reward $R(s)$ by binary discriminator and a distance learned by InfoNCE based on reachability

- States adjacent in a trajectory should have close embeddings, and vice versa

Solve in the Lagrange dual space:

- Single-level convex optimization (logsumexp+linear) with Fenchel dual over Lagrange dual variables λ
- Theoretical guarantee: equivalent to SMODICE with certain choice of coefficient for regularizers, and distance $c(s_i, s_j)$ that is independent of s_j

Learning Policy: weighted behavior cloning

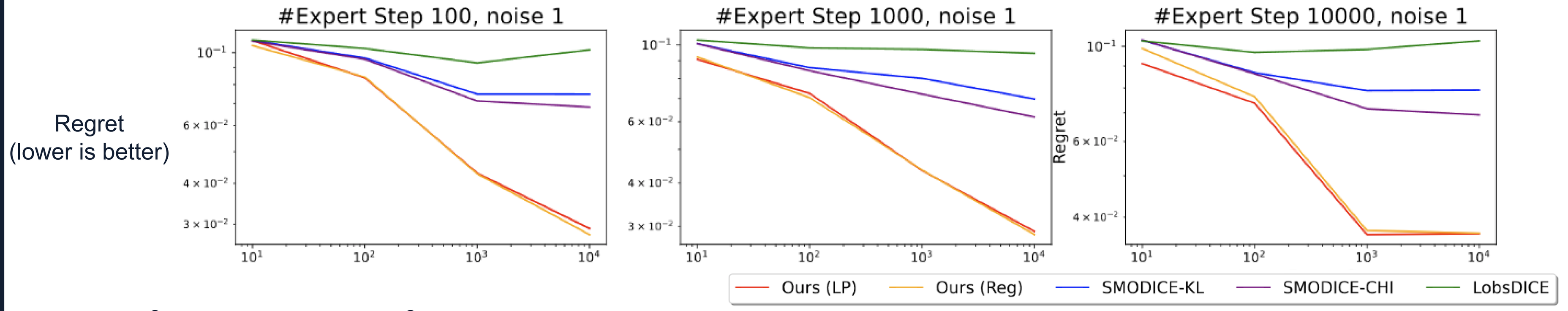
- weight $w(s_i, a_j, s_k) \propto \exp(\text{linear function of } \lambda)$

Results

Tabular MDP

Optimized with CVXPY

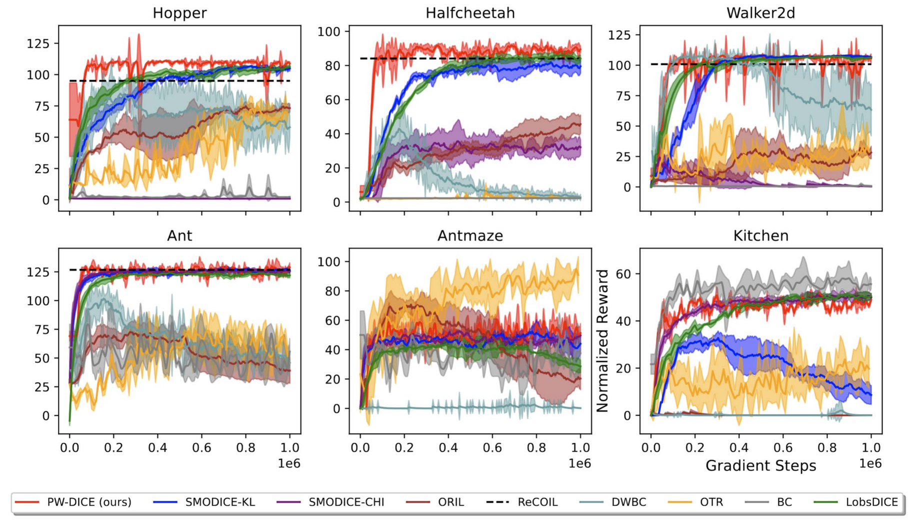
Ours with(out) regularizer illustrated in red(orange)



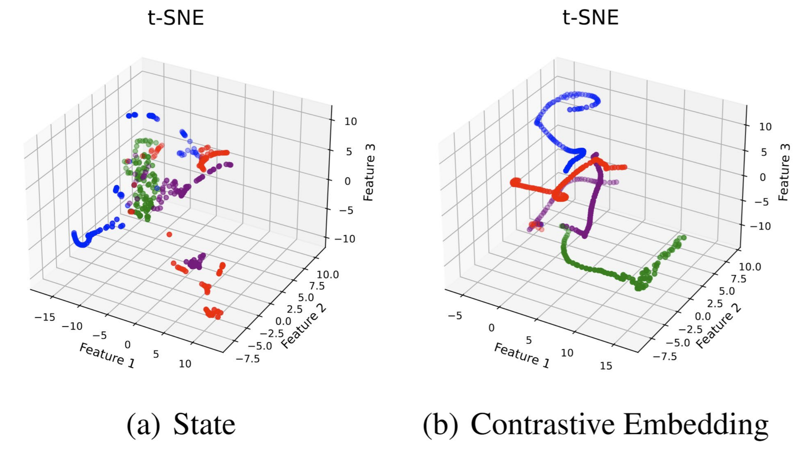
Mujoco Environments

Optimized with neural network

Our results illustrated in red outperform baselines and bring state embeddings in the same trajectories close



normalized reward (higher is better)



T-SNE of states and their embedding

Conclusion

- **Our key contribution:**
  - Shed light on importance of used distance metric
  - A novel LfO method generalizing, outperforming and removing theoretical assumption [1] of prior work
- **Limitation:**
  - Biased estimation of logsumexp in objective