Offline Imitation from Observation via Primal Wasserstein State Occupancy Matching



Kai Yan, Alexander G. Schwing, Yu-Xiong Wang
University of Illinois Urbana-Champaign

Why offline imitation from observations?

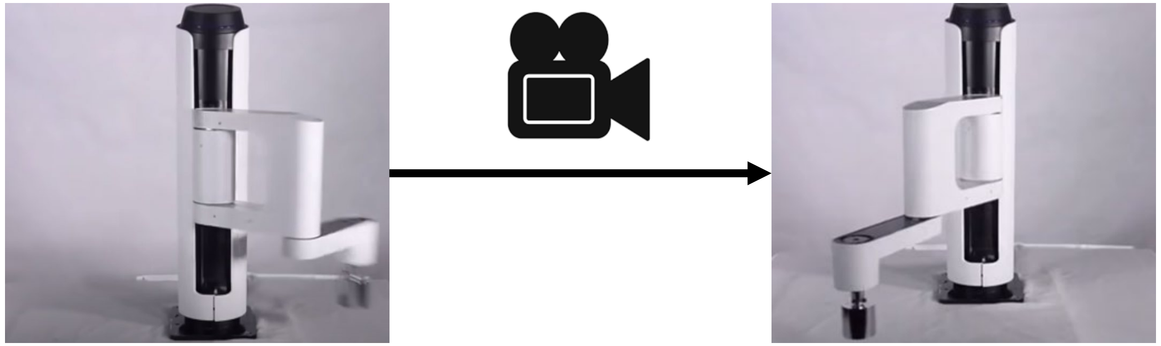
- Offline Imitation Learning (IL) is efficient
  + No reward labels

  + No online interactions

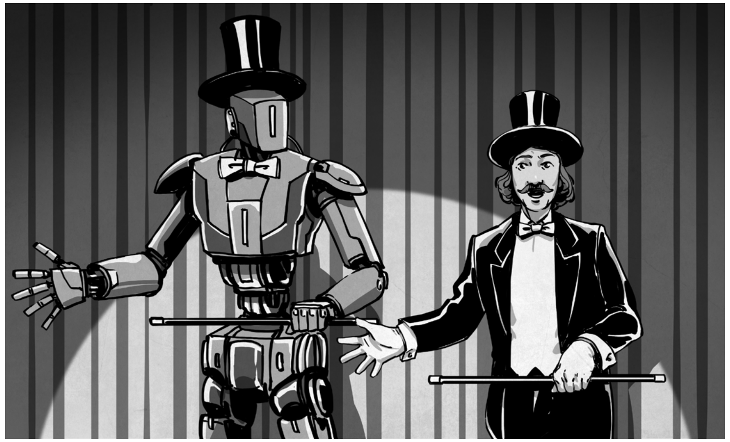
  – Requires expensive expert demonstrations which are hard to obtain

Why offline imitation from observations?

- Offline Imitation Learning (IL) is efficient, but expert demonstrations are few and sometimes state-only



Learning from video



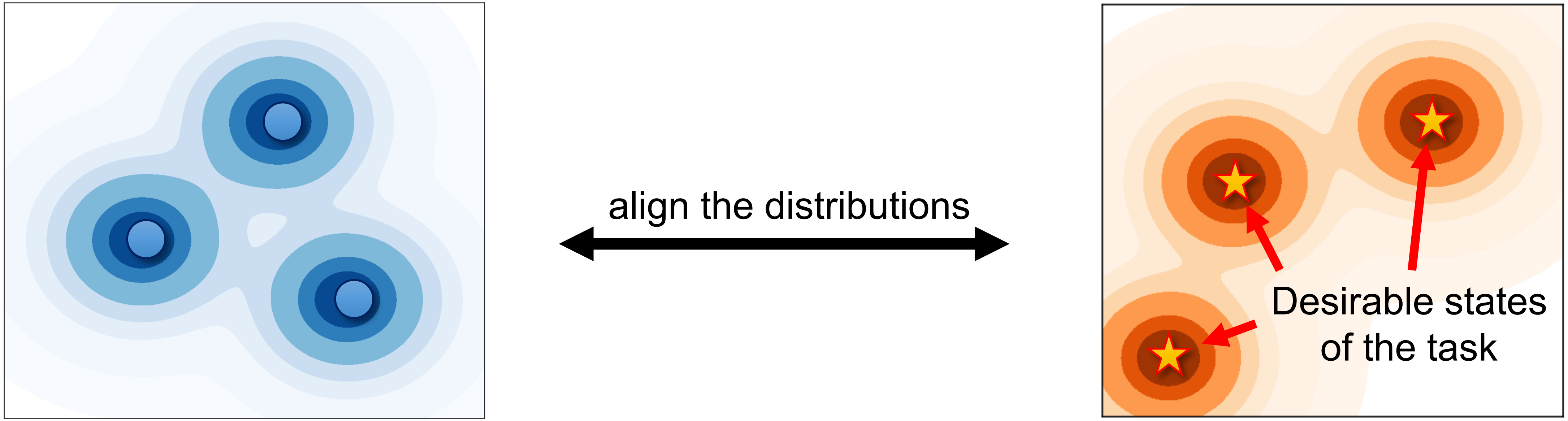
Embodiment difference

- Learn from few expert states + large, mixed-quality state-action dataset

Image Source: Internet

State occupancy matching

- How should we mimic the expert with its actions unknown?
  - Make state distribution (occupancy) between the learner’s and the expert’s policy close!

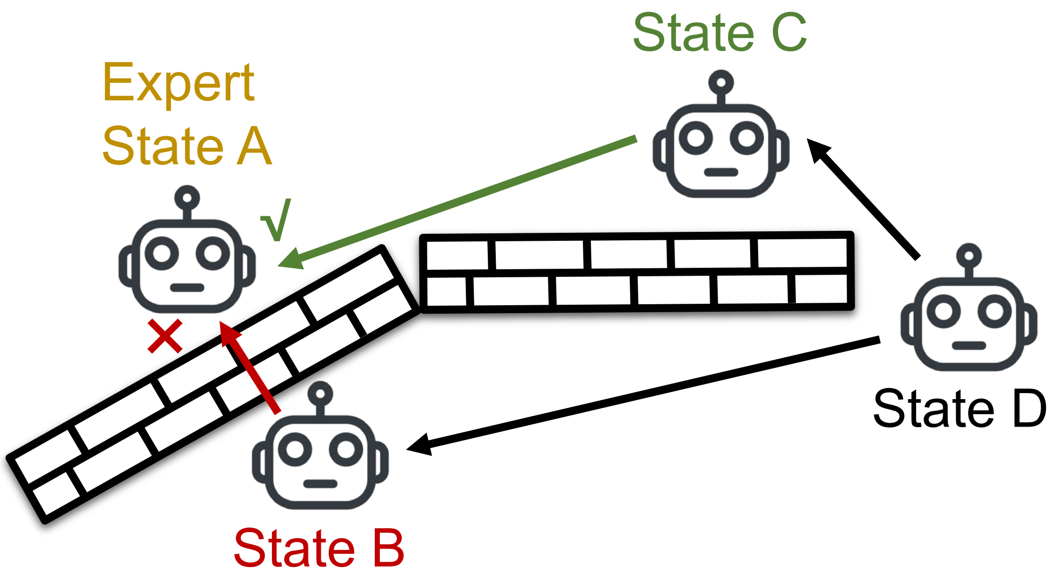


Learner states and distribution

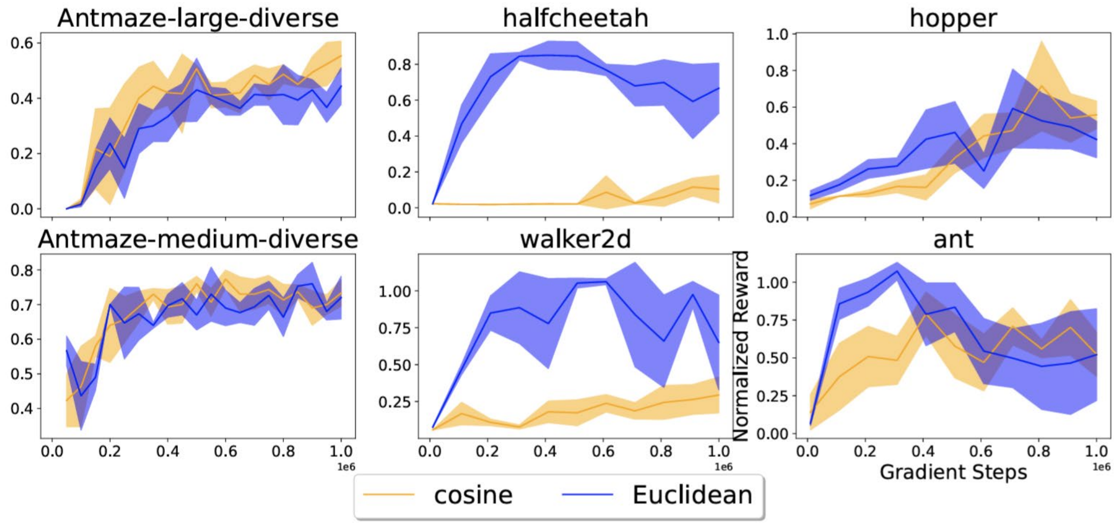
Expert states and distribution

Distance metric matters

- How should we define "close"?
  - State B or C – which is closer to A?
  - f-divergences (e.g. KL) cannot grasp the underlying geometric property between states
  - Wasserstein distance might help – but it depends on the underlying metric



OTR [1] with different distance metrics

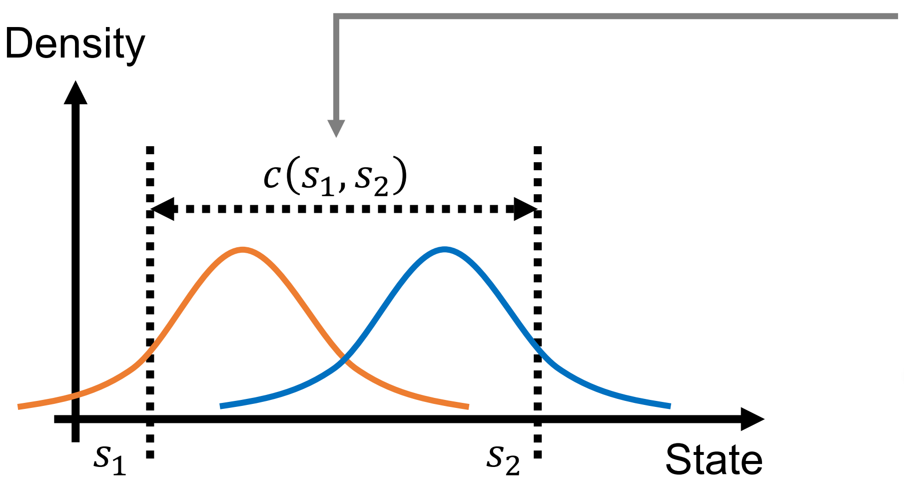


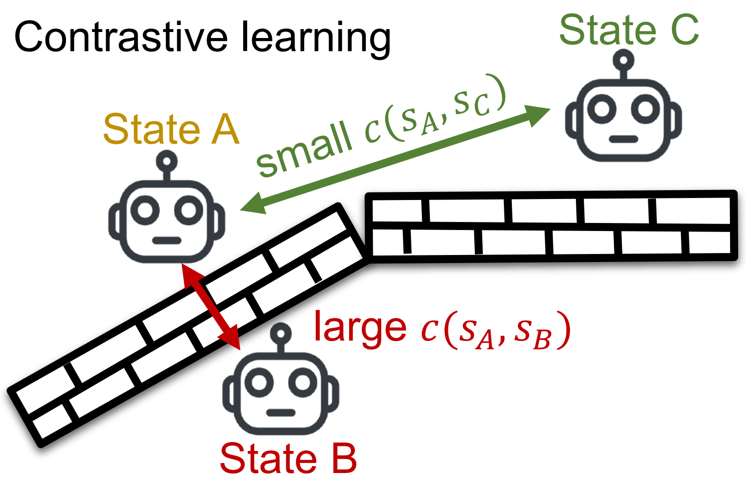
- We want to make the distance metric flexible, then learn a good one!

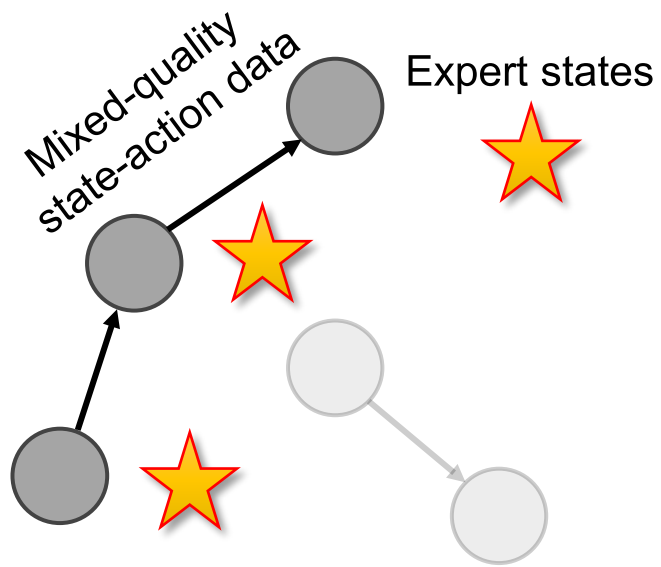
[1] Y. Luo et al. Optimal transport for offline imitation learning. In ICLR, 2023.

Our solution: primal Wasserstein + contrastive metric

- Primal Wasserstein distance allows using customized distance metric $c(s_1, s_2)$
  - With pessimistic regularizers, becomes a single-level unconstrained optimization
  - SMODICE [1] is a special case of our method with certain metric and hyperparameters
- Contrastive distance metric captures "reachability" in the dataset
- Weighted behavior cloning retrieves the learner's policy



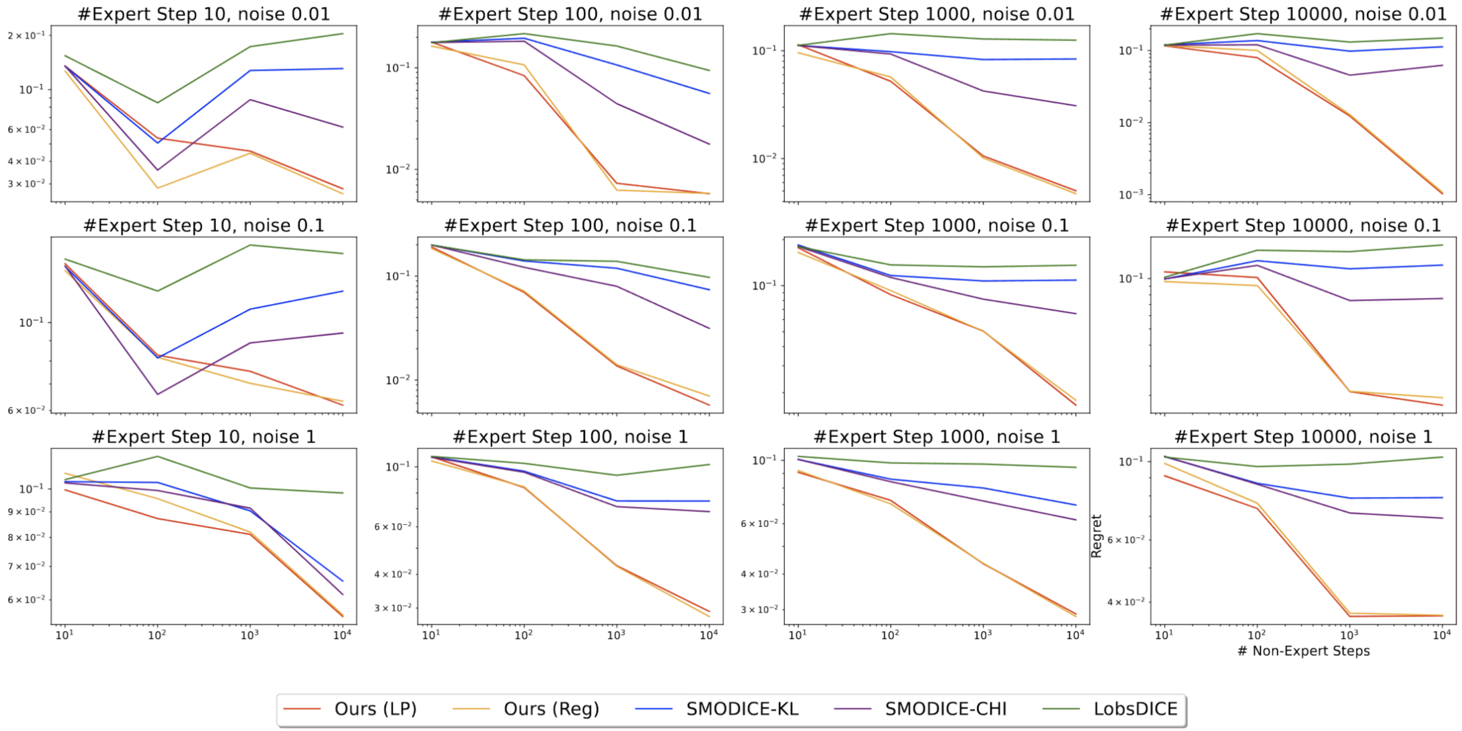




[1] Y. J. Ma et al. SMODICE: Versatile offline imitation learning via state occupancy matching. In ICML, 2022.

How well does our solution work?

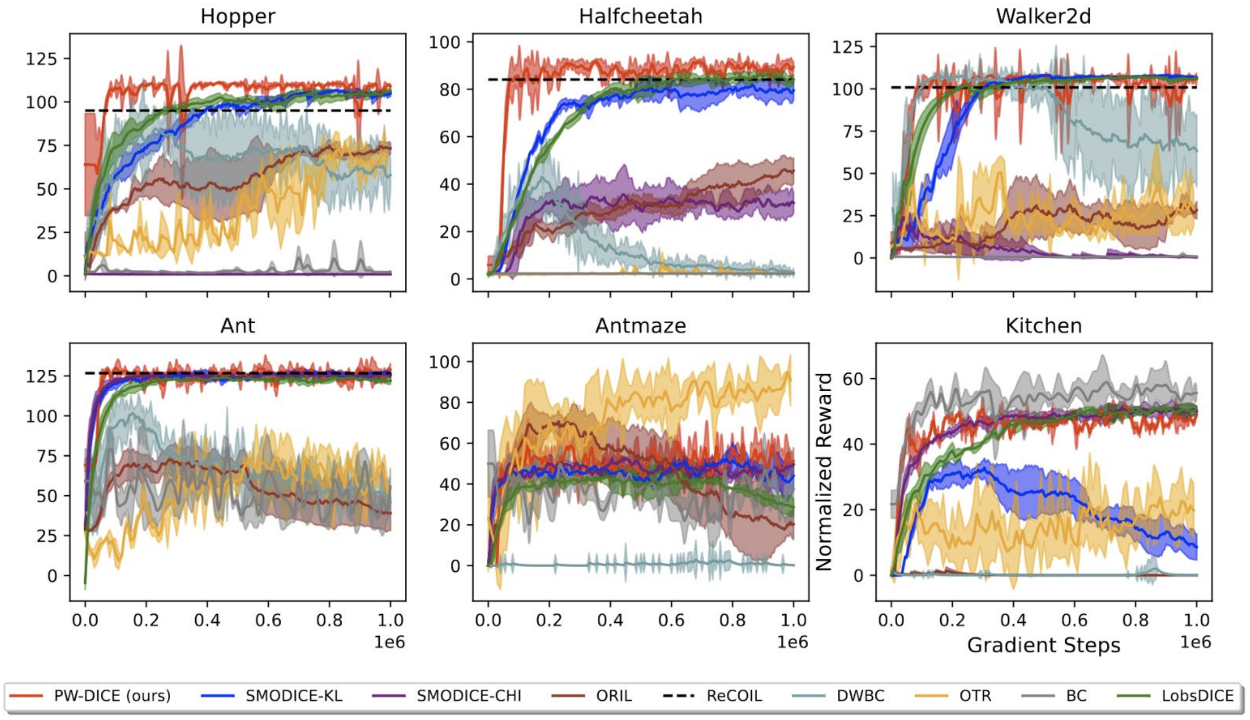
- Achieves lower regret on tabular MDP under many different settings
  - We test various dataset sizes and environment noise levels
  - Ours (red without regularizer / orange with regularizer) prevails consistently



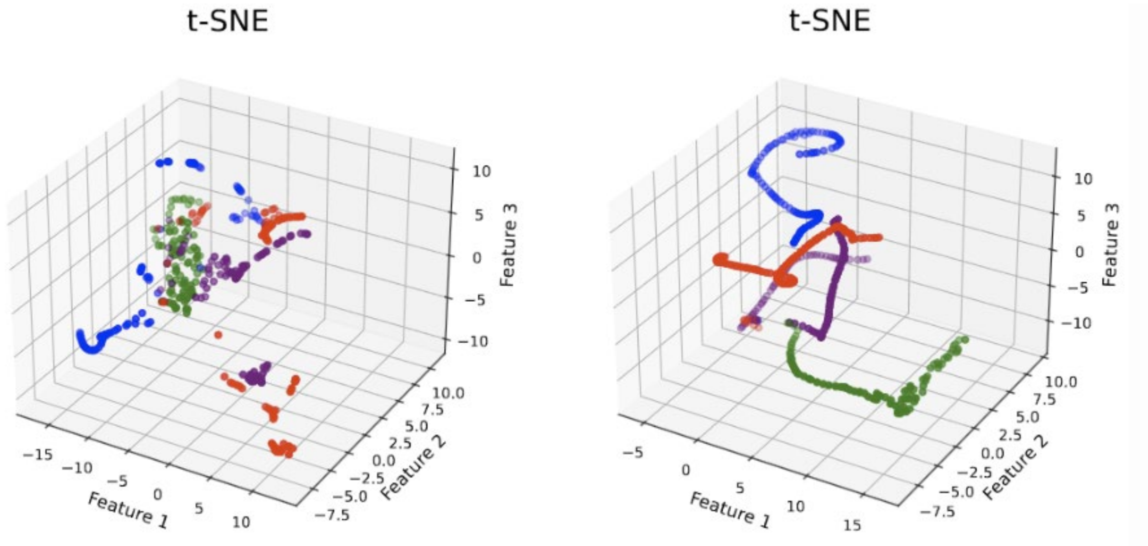
Regret (lower is better)

How well does our solution work?

- Our solution, with learned metric, also works well in continuous cases



Average reward (higher is better); ours highlighted in red



Trajectories in original state space

Trajectories in embedding space

t-SNE shows embedding of our learned metric better grasps the reachability of the states in trajectories

Thank you!

Feel free to contact kaiyan3@illinois.edu for any question!

Code repository

https://github.com/KaiYan289/PW-DICE

Website

https://t.ly/yKi9V

PDF of our paper

arXiv: 2311.01331