Hi I’m Kai Yan, and today I’ll tell you about our paper “**Offline Imitation from Observation via Primal Wasserstein State Occupancy Matching**”.

We will explain our work by first examining the problem we are trying to solve, which is offline imitation from observations. Why do we need offline imitation from observations? Different from online reinforcement learning (RL), offline imitation learning (IL) removes both the need for reward labels, which are not always available, as well as online interactions, which could be costly and often many are needed. However, offline IL has its own shortcomings: it needs expensive expert demonstrations for every new task.

To address this shortcoming, in this paper we focus on the expert data shortage where expert demonstrations are state-only. This is the case when trying to learn from video data, where the expert action is unavailable, or when training a robot from another embodiment, where the expert action cannot be transferred. In both scenarios, we want to Learn from a few expert states and a larger unlabeled, mixed-quality state-action dataset.

With the objective stated, the next problem is obvious: how can one mimic the expert with its actions unknown? To do so, we adopt a tried-and-true imitation learning framework called occupancy matching. It aims to make the state distribution between the learner’s and the expert’s policy close to each other. This method is based on the intuition that the expert is trying to reach desirable states regarding the task of interest: the learner should visit the desirable states favored by the expert and avoid the states that the expert never visits. By doing so, the two state distributions in the state space, which is called occupancy, are as close to each other as possible.

But with this idea, another vital problem emerges: what’s the definition of “close”? To see why this is important, consider a simple navigation task where the robot tries to go from state D to some expert state A, and it reaches either state B or C. The problem is, which should be considered closer to A? a popular choice is to use an f-divergence as the metric. However, since the state space can be large, if the expert seldomly visits either state B or C, then an f-divergence cannot help because it cannot grasp the underlying geometric property between the states. Thus, a few prior works turn to the Wasserstein distance; however, those works use only simple metrics such as the Euclidean distance, which is apparently misleading in this case as state B is closer to A in Euclidean distance but apparently further from A in geodesic distance. We further empirically verified the importance of the metric by testing OTR, a popular Wasserstein imitation learning algorithm with different distance metrics; the result shows that performance varies a lot with different distance metrics. This observation leads to the core inspiration of our paper: we want to make the distance metric flexible, and then automatically learn a good one.

So, how do we achieve this goal? In our paper, we propose to optimize the primal Wasserstein distance with a learned, contrastive distance metric. Primal Wasserstein distance minimization allows us to use a customized distance metric. It can be converted into a single-level unconstrained optimization with pessimistic regularizers that encourages the agent to stay in the area supported by data in continuous state space. We further showed that SMODICE, an established imitation learning method, is a special case of our method with a certain metric and hyperparameters. For the distance metric, we propose a contrastive distance metric that captures “reachability” in the dataset; it is based on learned embeddings, which attracts adjacent states in the trajectory of the mixed dataset and repels other states to address the robot dilemma we analyzed here. Finally, we use a weighted behavior cloning to retrieve the learner’s policy from the computed occupancy, where expert-related state-action pairs in the mixed-quality dataset are given more weight and vice versa, as implied by the transparency difference.

How well does this approach work? We have conducted extensive experiments on tabular MDPs with many different settings of dataset sizes and noise levels. The results show that our method, either with or without the regularizer for continuous MDPs, outperforms the baselines consistently.

And with our proposed contrastively learned metric and the regularized objective, our solution also works well in continuous cases. This is shown on this slide. We test our method on several mujoco environments and compare to many baselines using average reward. It is shown that our method, highlighted in red, generally works very well. We also visualize the t-SNE embedding of the trajectories in the original state space and the space with the embedding of our learned metric; we observe that the embedding space of our distance metric connects the states in the same trajectory that are apart in the original input, being aware of the reachability between the states.

Please scan the QR code for more information about our paper. Thanks for listening and don’t hesitate to reach out if you have any questions!