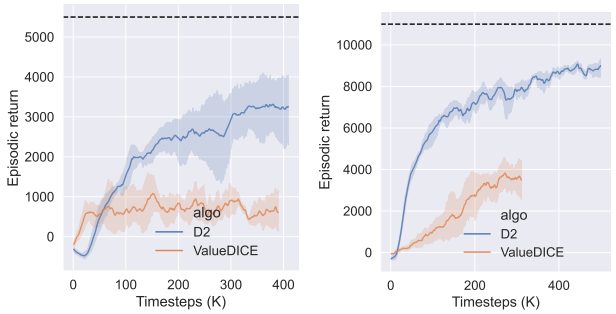
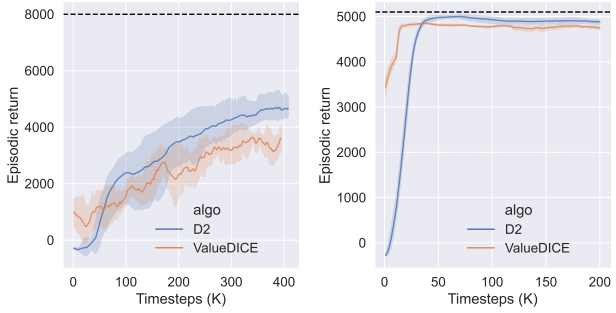
Figure 3: Performance comparisons using different numbers of demonstration trajectories (5, 10, 15, and 20); y-axis: normalized episodic returns after training for *100K* environment interactions (1.0 indicates the returns of an expert policy).



(a) DDPG demonstrations (b) SAC demonstrations



(c) TD3 demonstrations (d) TRPO demonstrations

Figure 4: Comparison with ValueDICE using demonstrations generated from different algorithms. The expert performance is also indicated by the black dashed line.

	IP	MCC	BW	LLC
	Returns for 50K training timesteps			
Expert PPO	1000.0	-0.050	302.49	181.21
BC	134.2	-0.159	-106.30	-75.62
DAC	**1000.0**	-0.085	297.58	-73.49
D2	**1000.0**	**-0.045****	**308.32**	**-34.63****
	Wall clock time for 50K timesteps (s)			
DAC	2755.0	2937.4	2458.8	2708.4
D2	**1526.9****	**1540.9****	**1720.2****	**1586.2****

Table 1: D2 vs DAC on InvertedPendulum (IP), MountainCarContinuous (MCC), BipedalWalker (BW) and LunarLanderContinuous (LLC); $**$ for significance (t-test).

ing performance of two algorithms on the HalfCheetah environment (we pick HalfCheetah as it was reported that learning the expert policy could take more than 20 million timesteps for imitation learning methods (**?**)). Although ValueDICE achieves good sample efficiency with demonstrations given by TRPO (stochastic policies), its performance becomes worse when demonstrations change. Empirically, ValueDICE is sensitive to demonstrations, and can fail to learn the expert policy. For example, ValueDICE initially progresses on DDPG demonstrations, then plateaus after 50K time-steps and never reaches expert performance. Furthermore, we found that even with demonstrations that are generated by stochastic policies, e.g., SAC, D2 still outperforms ValueDICE. Intuitively, this could be because a stochastic policy can always be turned into deterministic by picking the best or mean action, the discriminator in D2 is trained to predict that action for each state, and off-policy TD learns a deterministic policy that distills such action information from replay buffers. Overall, compared with ValueDICE, the performance of D2 imitation is more consistent across different types of demonstrations.

Ablation Studies

The discriminator is crucial for D2-Imitation to work properly and guarantees convergence to expert performance. We perform ablations on the discriminator and compare the

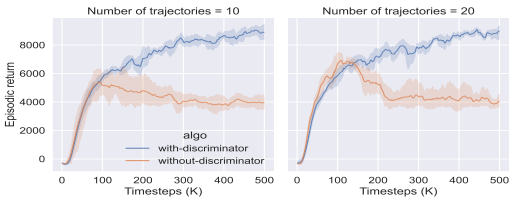


Figure 5: Effect of the discriminator on the training performance with 20 and 10 SAC demonstration trajectories.

training performance of D2-Imitation with one variant: one without any discriminator (denoted *without-discriminator*), which just puts on-policy samples to $\mathcal{B}^0$, assigns 0 reward to them and gives $+1$ reward to demonstration samples in $\mathcal{B}^+$, an idea adopted in Soft-Q Imitation Learning (SQIL) (**?**). Figure 5 shows that training quickly plateaus or even collapses if no discriminator is applied. Furthermore, this plateauing effect happens more quickly when fewer demonstration trajectories are used. This phenomenon has also been reported in SQIL. To avoid such training collapse, SQIL requires early stopping of the training process by judging whether the squared soft Bellman error converges to a minimum, which we argue can be challenging as this early stopping also relates to the number of trajectories used in training. Moreover, even with perfect early stopping (say 100K interactions as in the given figure in which the true reward function is known), the policy still fails to achieve the expert performance.

Related Work

Adversarial imitation casts imitation learning as a distribution matching problem (**?**) and leverages GANs (**?**) to minimize the Jensen-Shannon divergence between distributions induced by the expert and the learning policy. This approach avoids the difficulty of learning reward functions directly from demonstrations but is generally sample intensive. To improve sample efficiency, many methods extend adversarial imitation to be off-policy. For instance, Discriminator-actor-critic (DAC) (**??**) improves sample efficiency by reusing previous samples stored in a relay buffer. However, this approach still relies on non-stationary reward signals generated by the discriminator, which can make the critic estimation hard and training unstable. Recent work proposes to train a fixed reward function by estimating the support of demonstrations and training critics with the fixed reward (**?**). This support estimation itself could be hard given that only a limited number of empirical samples are available from the considered distributions. Another line of off-policy distribution matching approaches focuses on estimating the critics directly without learning any reinforcement signals (**??**). The state-of-the-art along this line is ValueDICE (**?**), which casts distribution matching as off-policy density ratio estimation and updates the policy directly via a max-min optimization. However, as we show in the analysis and experiments, these methods can be ill-posed when the demonstrations are generated from deterministic policies, and their performance can also be sensitive to the demonstrations used in training.

Recently, some non-adversarial imitation learning approaches have been proposed. For example, offline non-adversarial imitation learning (**?**) reduces the min-max in ValueDICE to policy iteration, which however still requires estimating density ratios and can potentially inherit the same issues from ValueDICE. By contrast, D2-Imitation avoids density ratio estimation and is more robust to different demonstrations. Primal Wasserstein imitation learning (**?**) avoids adversarial training by introducing an off-line cost function, which, however, requires a domain-specific metric and can be challenging to properly specify for different environments. D2-Imitation does not need such domain knowledge and can be applied in many different settings. Our reward design for D2-Imitation looks superficially similar to that of Soft Q Imitation Learning (SQIL), which assigns $+1$ reward to demonstration and zero for all interaction samples (**?**). However, D2-Imitation is fundamentally different in that the reward assignment is theoretically consistent with the use of deterministic policies.

Conclusion

In this paper, we revisited the foundations of adversarial imitation. We leveraged the similarity between the Bellman equation and the stationarity equation to derive a TD learning approach, which directly learns a special proxy, i.e., basic feasible solutions, for the expert state-action distribution. Moreover, we showed that the use of deterministic policies simplifies TD learning and yields a practical learning algorithm, D2-Imitation, which operates by first partitioning samples into two replay buffers and then learning a deterministic policy via off-policy deterministic policy gradients. Finally, the notion of partitioning samples into two groups theoretically follows from the use of a deterministic policy. Our empirical results demonstrated that D2-Imitation is effective in achieving good sample efficiency, and outperforms many adversarial imitation approaches on different control benchmark tasks with demonstrations generated by either deterministic or stochastic policies. Also, D2-Imitation consistently outperforms the state-of-the-art off-policy distribution matching method when training with various different types of demonstrations. In conclusion, D2-Imitation, as a direct result of leveraging two novel insights in the distribution matching formulation, is a simple yet very effective sample-efficient imitation learning approach.

Acknowledgements

Mingfei Sun is partially supported by funding from Microsoft Research. The experiments were made possible by a generous equipment grant from NVIDIA.

Aspernatur iste optio molestias suscipit rerum eius magnam aperiam error, error sed nisi quam architecto corrupti itaque, numquam officiis laborum facere incidunt quis asperiores deserunt, aspernatur illum cum ipsa natus?Id vero nesciunt repellendus soluta facilis laudantium corrupti expedita, accusantium consequatur libero aut id delectus ea, asperiores saepe cum hic sunt, perferendis at repellendus corrupti sapiente fuga nihil ipsa.