Sample Efficient Imitation Learning Through Transfer

Michael Kelly (mkelly2@stanford.edu)

Motivation

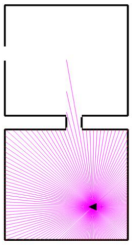
- Interactive imitation learning methods such as DAGGER address the compounding error issues that plague behavioral cloning [1]
- However, they require the continual presence of an expert throughout training, which is a limitation in domains where access to the expert is expensive, such as in human-in-the-loop imitation learning [2][3]
- This work focuses on optimizing the sample complexity of DAGGER and related algorithms

Method

- Assumes access to an imperfect simulator of the target environment
- Perform reinforcement learning (e.g. TRPO) to train a policy in the simulator; use this policy as an initialization for the novice policy in DAGGER

Experiments

- Agent is a Dubins car initialized with a random pose in the lower room; must pass through the narrow passage to the exit in the upper room
- State transitions are affected by zero-mean Gaussian noise and a persistent drift force
- Observations are noisy lidar scans of the environment
- Simulators capture only a portion of the drift force and none of the observation noise



References

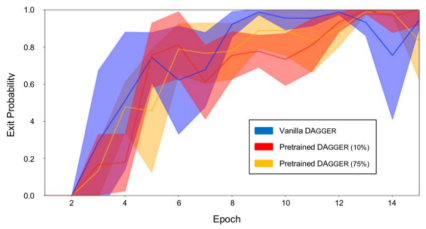
[1] Ross, S.; Gordon, G. J.; and Bagnell, J. A. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics*, 627–635.

[2] Laskey, M.; Staszak, S.; Hsieh, W. Y.-S.; Mahler, J.; Pokorny, F. T.; Dragan, A. D.; and Goldberg,K. 2016. Shiv: Reducing supervisor burden in dagger using support vectors for efficient learning from demonstrations in high dimensional state spaces. In *ICRA*, 462–469.

[3] Zhang, J., and Cho, K. 2017. Query-efficient imitation learning for end-to-end autonomous driving. In *AAAI Conference on Artificial Intelligence*.

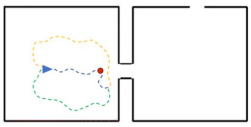
Initial results

- Chart shows the baseline performance of DAGGER, as well as its performance when the novice policy was pretrained in simulators where the drift force was 10% and 75% of that of the target environment
- Results:
  - No discernible benefit to pretraining in the higher vs. lower fidelity simulators
  - No discernible benefit to pretraining of any kind over the baseline approach, where the novice is initialized at random



Catastrophic Forgetting

- **Catastrophic Forgetting**: refers to the tendency of artificial neural networks to abruptly lose previously learned knowledge about a task when trained on new data relevant to a different task
- Likely exist various distinct policies with similar levels of performance, many of which are reachable with TRPO



- Because of this diversity of good policies, when performing imitation learning we will often be switching between highly distinct behavioral modes, leading to catastrophic forgetting
- Solution: **Multitask Learning**
  - Train simultaneously rather than sequentially
  - In this case: **interleave rounds of behavioral cloning into the RL pretraining phase**
  - Appears to minimize forgetting and improve learning performance; novice can be fine-tuned rather than trained from scratch
- Future work
  - Incorporate elastic weight consolidation, employ finer interleaving of RL and IL, more extensive testing