Efficient Imitation Learning with Local Trajectory Optimization

Jialin Song, Joe Wenjie Jiang, Amir Yazdanbakhsh, Ebrahim Songhori, Anna Goldie, Navdeep Jaitly, Azalia Mirhoseini
Imitation Learning

• Learning from expert demonstrations.

• It can be more sample efficient than RL, especially in sparse reward environments.

• The convergence speed of learning depends on how expert demonstrations are collected.
How to Collect Demonstrations?

Interpolate between DAgger and BC
Local Trajectory Improvement
Theoretical Justification

\[ J(\pi) \geq \sum_{i=1}^{T/t} \gamma^{t(i-1)} J_{\pi^{t:i+1}}(\pi^*) - \frac{1 - \gamma^T}{1 - \gamma^t} t^2 \varepsilon \]

\( \varepsilon \) measures how often \( \pi \) and \( \pi^* \) disagrees

The quality of a policy \( \pi \)
Finding the Balance

\[ J(\pi) \geq \sum_{i=1}^{T/t} \gamma^{t(i-1)} J_{\pi_i:t_i+1}(\pi^*) \left( \frac{1 - \gamma^T}{1 - \gamma^t} t^2 \epsilon \right) \]

- Both terms are monotonic increasing functions in t
- Find a value of t between 1 and T to maximize the RHS
Experiment Setup

• MuJoCo Control Environment.
  • Each trajectory has a max time horizon of 1000.

• $\pi^*$: use Monte-Carlo tree search with a current policy $\pi$.
  • Similar to the approach used in AlphaGo.

• Reference implementation: https://github.com/google-research/google-research/tree/master/polish
Experiment Results: Compare with Baselines

- An intermediate value of $t = 32$ outperforms both DAgger ($t=1$) and BC ($t=1000$).
- It also outperforms the PPO RL baseline.
Experiment Results: Parallelization Speedup

- The time to collect expert trajectories through MCTS does not increase too much when using a value of $t=32$. 
Thanks for your time!

Please find us in the virtual poster session if you have questions.