RL experiment - ReachingJaco-v1 with tuned hyperparameters

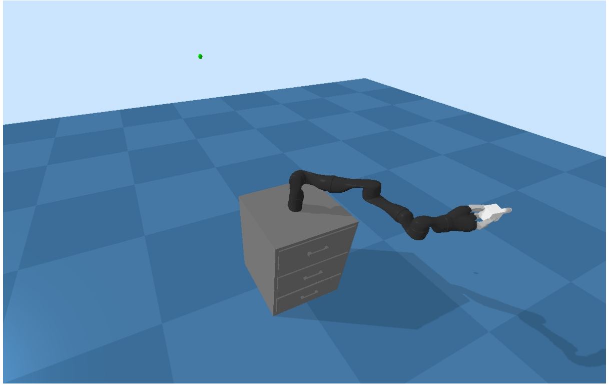

Figure 1: The ReachingJaco-v1 environment

1 Introduction and methods

The aim of this document is plot the results of the following RL experiment:

- Algorithms: ACKTR, PPO2, SAC, TD3
- Environment: ReachingJaco-v1
- Number of time steps: 2M
- Number of initialisation seeds: 2
- Number of parallel environments: 8 for ACKTR and PPO2 and 1 for SAC and TD3 (parallelisation not supported).

The performance metrics are defined as follows:

- Train time (min) : Wall time to train.
- Success ratio : number of successful episodes / number of reachable episodes
  An episode is successful if the distance between the finger tip and the target is less or equal to 0.03.
- Average reaching time : sum (number of time steps of all successful episodes) / number of successful episodes
  An episode has a maximum of 200 time steps.
- Efficiency: mean reward / mean training walltime.

2 Results

2.1 Raw results

2.2 Learning curves

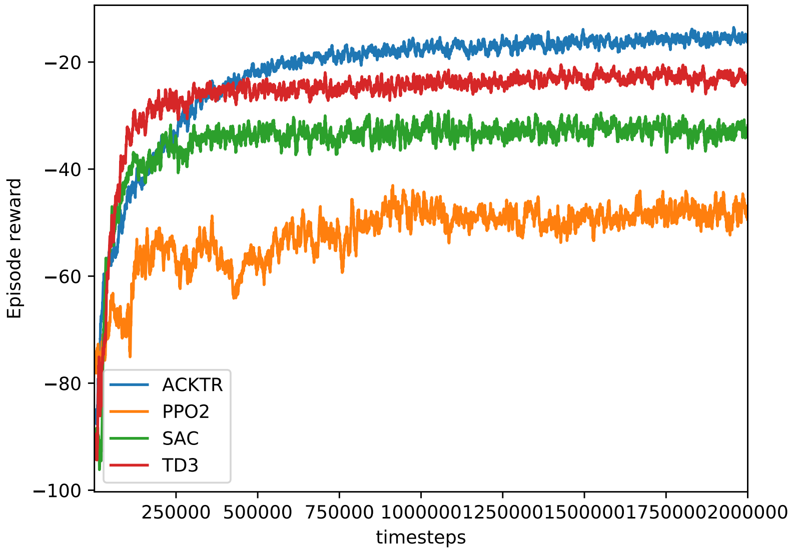

Figure 2: All learning curves.

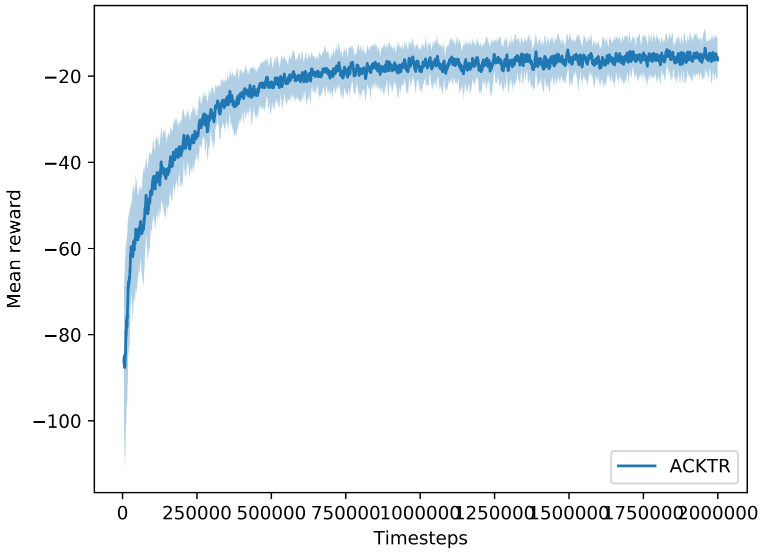

Figure 3: Learning curve ACKTR.

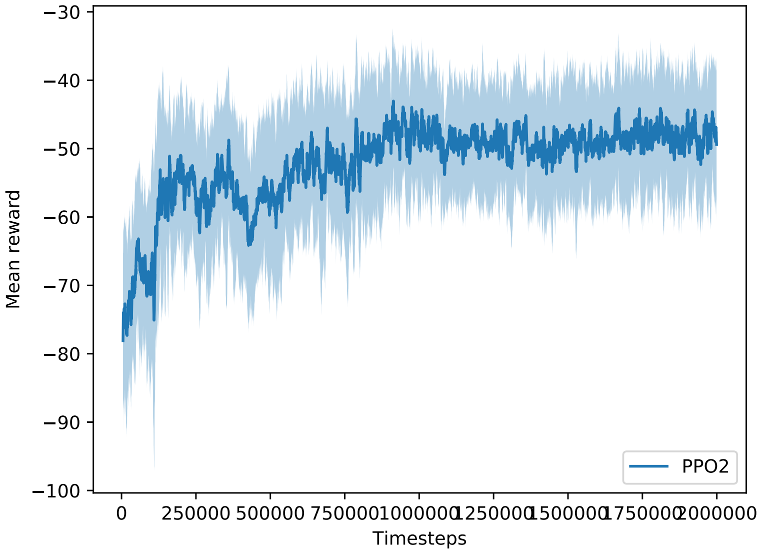

Figure 4: Learning curve PPO2.

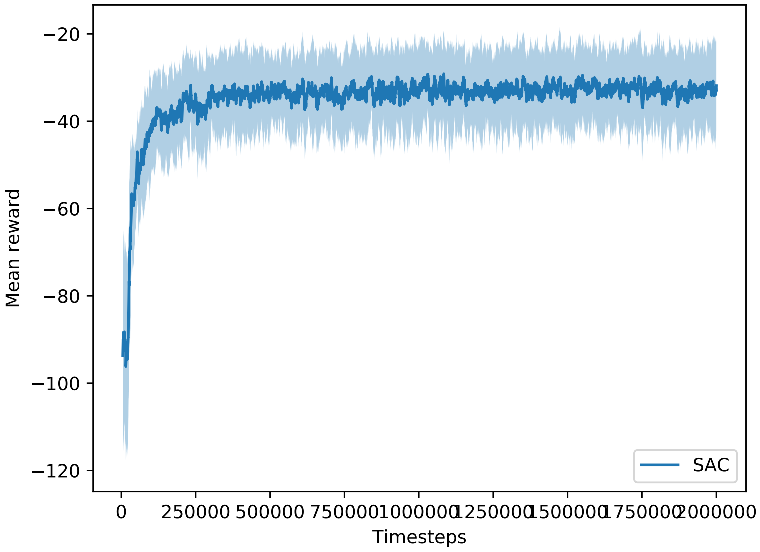

Figure 5: Learning curve SAC.

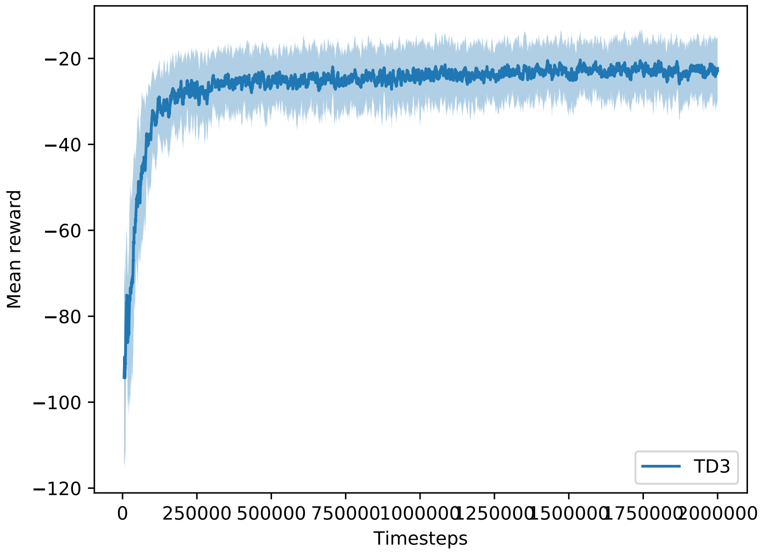

Figure 6: Learning curve TD3.

2.3 Evaluation

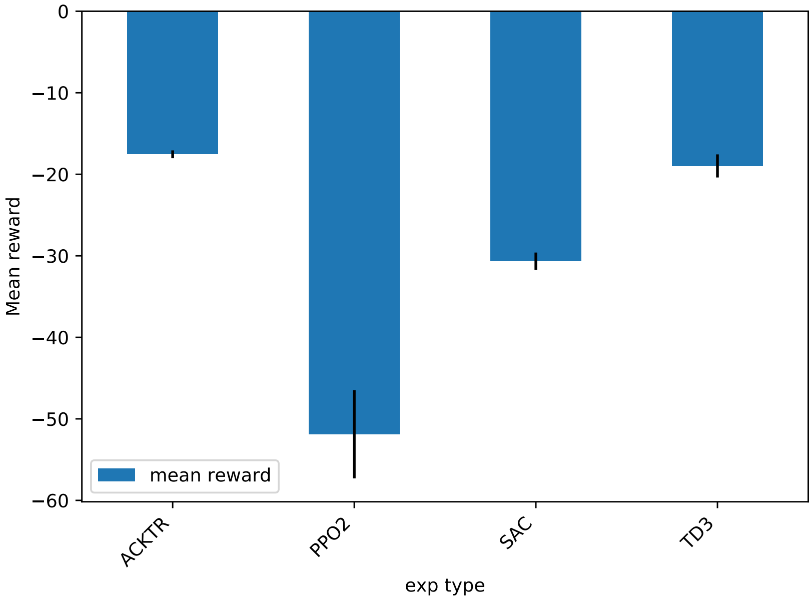

Figure 7: Reward vs algorithms.

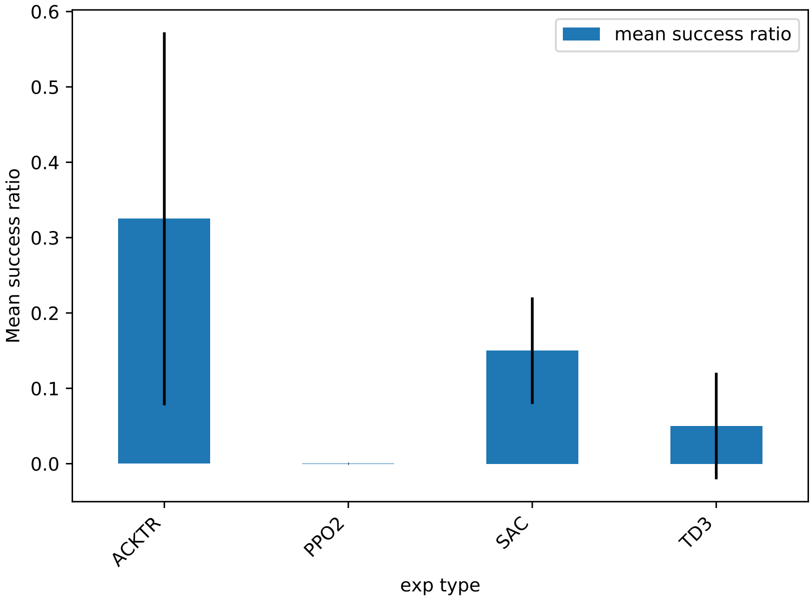

Figure 8: Success ratio vs algorithms.

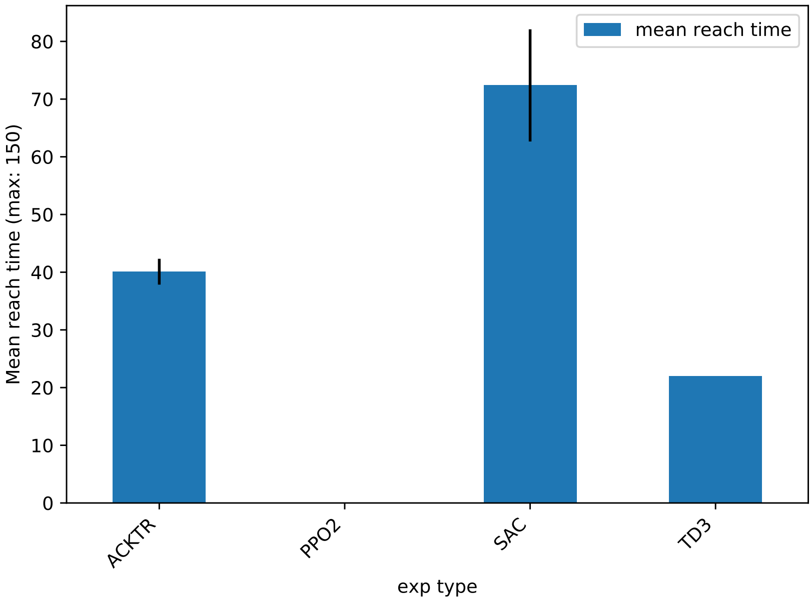

Figure 9: Reach time vs algorithms.

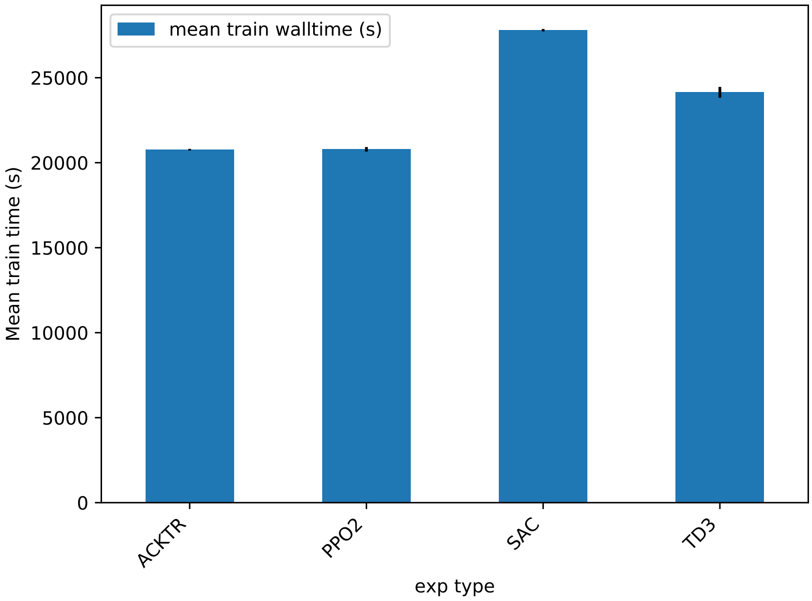

Figure 10: Train walltime vs algorithms.

3 Findings summary