Karlsruhe Institute of Technology

Autonomous Learning Robots Lab

Benchmarking Reinforcement Learning Algorithms on Realistic Simulated Environments

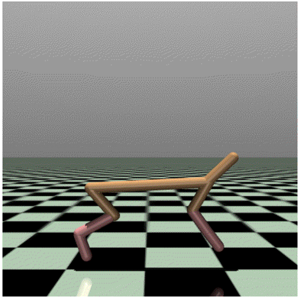
Bachelor’s Thesis

Jan Küblbeck | March 17, 2023

Contents

Motivation

- Benchmarking to test and compare new RL algorithms
- Robotics benefits from realistic simulations – ALR Simulation Framework
- Goal: Implement and apply a set of benchmark tasks using the Simulation Framework



Selected Algorithms

Proximal Policy Optimization (PPO)

- Policy gradient method with on-policy learning (Schulman et al. 2017)
- Using ratio of new and old policy & clipped objective to keep policy updates small

Selected Algorithms

Proximal Policy Optimization (PPO)

- Policy gradient method with on-policy learning (Schulman et al. 2017)
- Using ratio of new and old policy & clipped objective to keep policy updates small

Deep Deterministic Policy Gradient (DDPG)

- Off-policy actor-critic algorithm (Lillicrap et al. 2015)
- Uses experience replay
- Learns a Q-function and policy both using gradient ascent

Selected Algorithms

Karlsruhe Institute of Technology

Twin Delayed DDPG (TD3)

- Variation of DDPG with key modifications (Fujimoto, Hoof, and Meger 2018):
  - "Twin": learns two different Q-functions to limit overestimation
  - "Delayed": updates policy parameters less frequently than critic parameters

Selected Algorithms

Karlsruhe Institute of Technology

Twin Delayed DDPG (TD3)

- Variation of DDPG with key modifications (Fujimoto, Hoof, and Meger 2018):
  - "Twin": learns two different Q-functions to limit overestimation
  - "Delayed": updates policy parameters less frequently than critic parameters

Soft Actor-Critic (SAC)

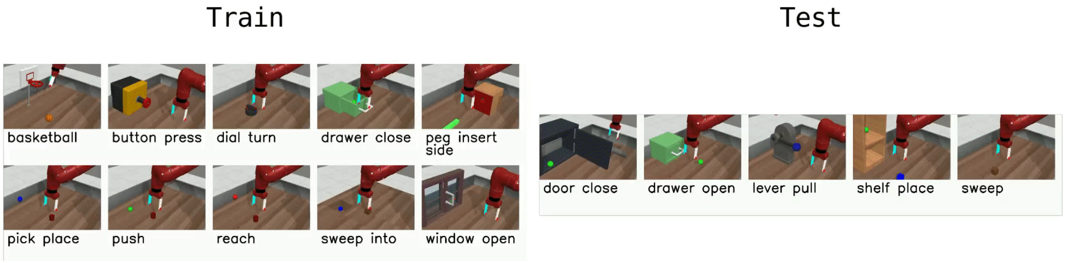
- Similar to DDPG/TD3, but learns a stochastic policy (Haarnoja et al. 2018)
- Aims to maximize policy entropy

How To Benchmark RL Algorithms

- Run many trials with different random seeds
- Report average performance instead of picking best-case results
- Use reliable implementations of benchmarked algorithms
- Equal hyperparameter tuning to avoid unfair advantages
- Use a comprehensive and diverse set of tasks

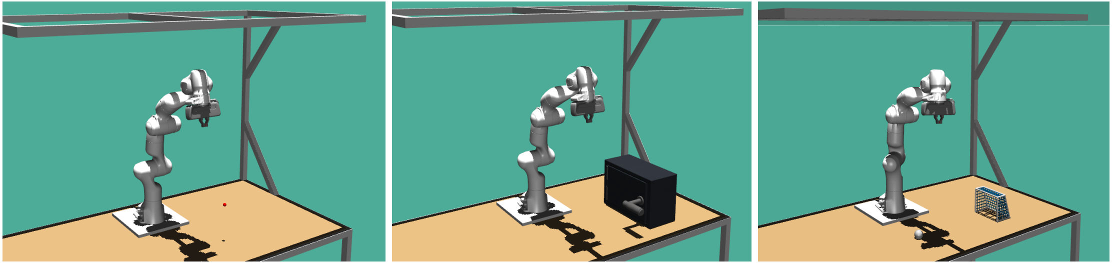
Related Work

- Abstract benchmarks: ALE, many gym environments
- RLBench: robot benchmark for RL and imitation learning (James et al. 2020)
- Meta-World: multi-task and meta-RL benchmark with 50 tasks (Yu et al. 2020)



Tasks

- Reach: Move end-effector to goal position. Goal location and initial robot position are randomized.
- Door Opening: Pull on a handle to open a door.
- Soccer: Move a ball into a goal. The ball's starting position is randomized.



Reach Task Implementation

Karlsruhe Institute of Technology

- Random sampling of initial robot position and goal location for every episode
- 34-dimensional observation space, including goal position g and end effector position p
- Step reward $r_t = -\exp(\|g - p\|^2)$
- Success threshold: $\|g - p\| < 0.025$, otherwise maximum episode length 250 steps

Door Opening Task Implementation

Karlsruhe Institute of Technology

- 47-dimensional observation space, including position of the door handle h and angle of the door θ
- Reward made of two components, which are multiplied with weights and added together:
  - $r_{handle} = ||p - h||$
  - $r_{hinge} = \exp(\theta - \frac{\pi}{2}) - 1$
- Episodic success when $\theta > \frac{\pi}{6}$ $(30°)$, maximum length 625 steps

Soccer Task Implementation

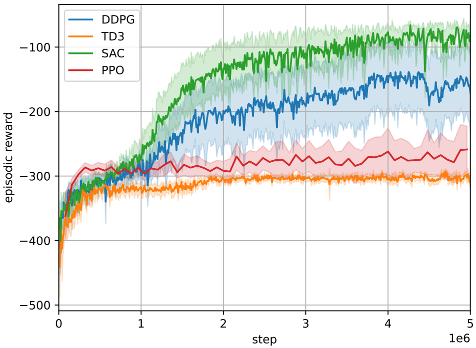
Karlsruhe Institute of Technology

- 50-dimensional observation space, including ball position b, goal location g and goal size
- Reward includes three components:
  - distance between end effector and ball $||p - b||$
  - distance between ball and goal $||b - g||$
  - constant penalty for missing the goal
- Success definition: Ball is in the goal; otherwise maximum episode length of 500 steps

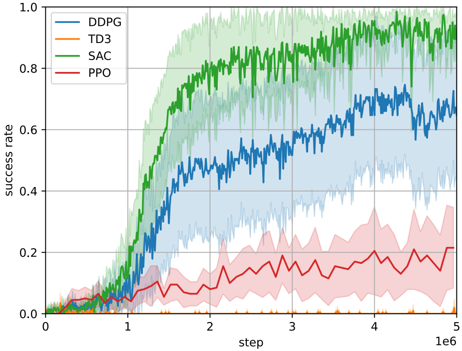
Experiments

- Trained for 5 million steps each
- Using 20 (reach, door opening) or 10 (soccer) different random seeds
- Benchmarking standard implementations of algorithms from Stable-Baselines3

Reach Task Results

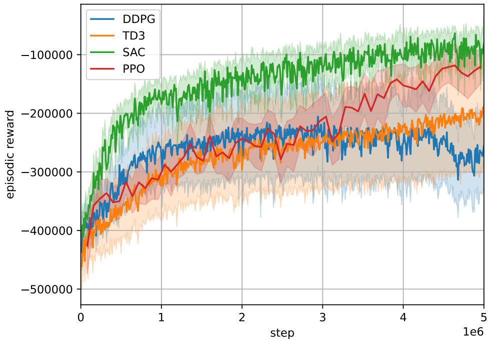


(a) Mean episodic reward

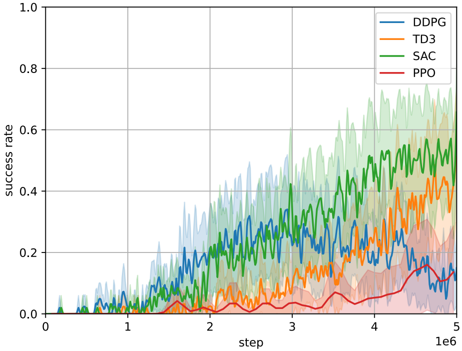


(b) Mean success rate

Door Opening Task Results

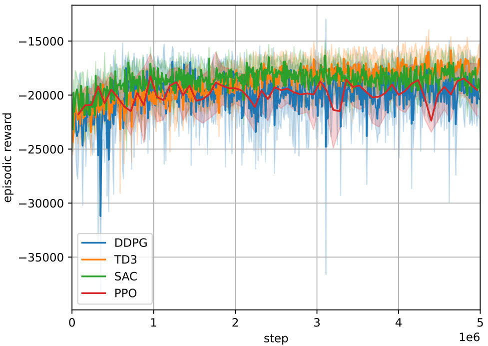


(c) Mean episodic reward

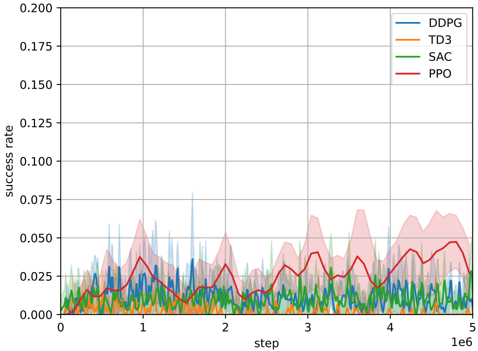


(d) Mean success rate

Soccer Task Results



(e) Mean episodic reward



(f) Mean success rate

Algorithm Evaluation

Karlsruhe Institute of Technology

- PPO
  - slow learning due to on-policy approach with low sample efficiency
  - best performer in very challenging soccer task
- DDPG
  - inconsistent performance between different seeds
  - capable of quick learning, but average success can degrade in the long term
- TD3
  - surprisingly unsuccessful in simple reaching task
  - outperformed regular DDPG in door opening task
- SAC
  - top performer in 2/3 tasks
  - can learn fast and consistently

Conclusion

- Introduced fundamentals of RL algorithms and benchmarking
- Implemented three tasks based on Meta-World
- Performed experiments and evaluated results with four algorithms

Future Works

- Improvements to the benchmark environments & development of additional environments using the framework
- Testing other algorithms on the benchmark tasks
- Sim-to-real transfer

Bibliography

[1] Scott Fujimoto, Herke Hoof, and David Meger. “Addressing function approximation error in actor-critic methods”. In: *International conference on machine learning*. PMLR. 2018, pp. 1587–1596.

[2] Tuomas Haarnoja et al. “Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor”. In: *International conference on machine learning*. PMLR. 2018, pp. 1861–1870.

[3] Stephen James et al. “Rlbench: The robot learning benchmark & learning environment”. In: *IEEE Robotics and Automation Letters* 5.2 (2020), pp. 3019–3026.

[4] Timothy P Lillicrap et al. “Continuous control with deep reinforcement learning”. In: *arXiv preprint arXiv:1509.02971* (2015).

[5] John Schulman et al. “Proximal policy optimization algorithms”. In: *arXiv preprint arXiv:1707.06347* (2017).

[6] Tianhe Yu et al. “Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning”. In: *Conference on robot learning*. PMLR. 2020, pp. 1094–1100.