A Motorized Wheelchair that Learns to Make its Way through a Crowd

Denis Steckelmacher, Hélène Plisnier, and Ann Nowé

Vrije Universiteit Brussel (VUB), Brussels, Belgium
dsteckel@ai.vub.ac.be
http://steckdenis.be/bnaic_demo_wheelchair.mp4

Abstract. Reinforcement Learning on physical robots is challenging, especially when the robot only provides high-dimensional sensor readings, and must learn a task in an environment for which no model is available. In this demonstration, a robotic wheelchair moves in a crowd of people. It is rewarded for moving fast, but punished when too close to a person or obstacle. Because the room the chair will be is unknown, and because it is impossible to model how every person moves, learning our *go-fast* task requires the use of model-free reinforcement learning. We propose to use Bootstrapped Dual Policy Iteration [4], a highly sample-efficient model-free RL algorithm, that we show needs only two hours to learn to avoid obstacles on our wheelchair.

1 Algorithm

Bootstrapped Dual Policy Iteration [4] is a model-free actor-critic reinforcement-learning algorithm for continuous states and discrete actions. Contrary to conventional actor-critic algorithms [3], BDPI's critic learns the optimal *off-policy* Q^* function with a variant of Q-Learning, instead of the *on-policy* Q^π function. BDPI's actor learning rule, based on Conservative Policy Iteration [2], is able to use *several off-policy* critics. The combination of highly sample-efficient off-policy critics, and a smart actor learning rule, makes BDPI one or two orders of magnitude more sample-efficient than state-of-the-art algorithms. Because the sample-efficiency of BDPI increases with the amount of computation done per time-step, we extend BDPI by processing many batches of experiences and many critics in parallel, *ala* Hogwild! [1], instead of sequentially training each critic on each set of experiences. This allows to fully utilize the 32 cores of an AMD Threadripper 2990WX machine, maximizing the *amount of learning* that happens per time-step.

2 Setting, Task and Safety

The reinforcement-learning agent running on the wheelchair is able to observe a 160-dimensional vector of distance readings, produced by a webcam pointing towards the front of the wheelchair, with a simple edge-detection algorithm. We assume a smooth

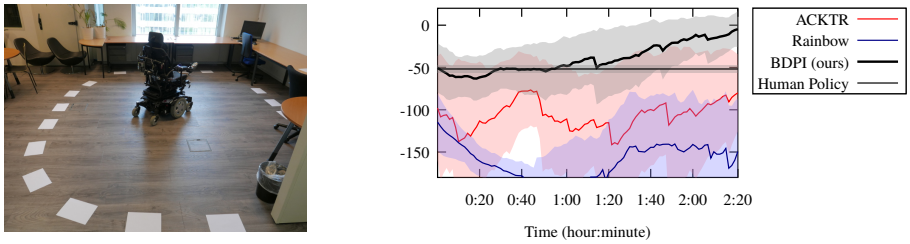

Fig. 1. *Left:* our motorized wheelchair in a 4-by-4 meters area. *Right:* BDPI is the only algorithm able to learn to avoid obstacles in 2 hours, and even outperforms a good human-designed policy. Our demonstration is a slightly more challenging task, as obstacles will be people, and the agent will be able to clear its throat to try to get people to make way for it.

floor, with few/weak edges on it, and observe that obstacles (shoes, people, walls) produces strong edges. The agent has access to four actions: going forwards, turning left, turning right, and emitting a slight coughing sound. Actions are executed every half-second. Over two hours, this leads to only 14K actions being executed, an immense sample-efficiency challenge for an agent observing a 160-dimensional continuous input. A positive reward is given whenever the agent goes forwards. A negative reward is given when the agent is too close to an obstacle (less than 40cm), or if any person presses the punish button on the wheelchair.

Because the wheelchair will move in the crowd of people looking at the demonstrations, safety is critical. The wheelchair will run at its slowest setting, a backup policy prevents the chair from going to close to an obstacle, the off-button of the wheelchair is easily accessible from all sides, and one of the authors will always be closely supervising the chair. Our demonstration requires one power plug and 1 square meter for our infrastructure. This demonstration can be contained in a 4×3 meters area, but would be more interesting if the chair is allowed to move amongst people.

Acknowledgments

The first and second authors are funded by the Science Foundation of Flanders (FWO, Belgium), respectively as 1129319N Aspirant, and 1SA6619N Applied Researcher.

References

1. Benjamin Recht, Christopher Re, Stephen Wright, and Feng Niu. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *Advances in neural information processing systems (NIPS)*, pages 693–701, 2011.
2. Bruno Scherrer. Approximate policy iteration schemes: A comparison. In *Proceedings of the 31th International Conference on Machine Learning (ICML)*, pages 1314–1322, 2014.
3. John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *Arxiv*, abs/1707.06347, 2017.
4. Denis Steckelmacher, Hélène Plisnier, Diederik M Roijers, and Ann Nowé. Sample-efficient model-free reinforcement learning with off-policy critics. *arXiv*, abs/1903.04193, 2019.