Sample-Efficient Model-Free Reinforcement Learning with Off-Policy Critics

Denis Steckelmacher, Hélène Plisnier, Diederik M. Roijers, and Ann Nowé

Vrije Universiteit Brussel (VUB), Brussels, Belgium
Contact: dsteckel@ai.vub.ac.be

1 Introduction

Sample efficiency is key to many applications of reinforcement learning in the real world, for instance when learning directly on a physical robot [1]. In discrete-action settings, value-based methods tend to be more sample-efficient than actor-critic ones [2]. We argue that this is because actor-critic algorithms learn a critic Q^{π}, that must accurately evaluate the actor, instead of Q^*, the optimal Q-function [3]. Some algorithms allow the agent to execute a policy different from the actor, which the authors refer to as off-policy, but the critic is still on-policy with regards to the actor [4,5]. We propose a new actor-critic algorithm, inspired from Conservative Policy Iteration [6], that uses *off-policy* critics that approximate Q^* instead of Q^{π}.

2 Bootstrapped Dual Policy Iteration

Our algorithm, fully described in [7], is divided in two parts: *off-policy* critics, and an *actor* that is robust to off-policy critics. Our critic learning rule, inspired by Clipped DQN [8], is given in Equation 1. This learning rule is used to train 16 critics, each of them on distinct 256-experiences batches sampled from a single shared experience buffer, as suggested by [9]. Our actor learning rule consists of, after every time-step, updating each critic i on a batch B_i of experiences, then sequentially updating the actor with Equation 2 with every batch $B_1 ... B_{16}$:

$$Q^{A,i}(s,a) \leftarrow Q^{A,i}(s,a) + \alpha\big(r + \gamma V(s') - Q^{A,i}(s,a)\big) \quad \forall (s,a,r,s') \in B_i \qquad (1)$$

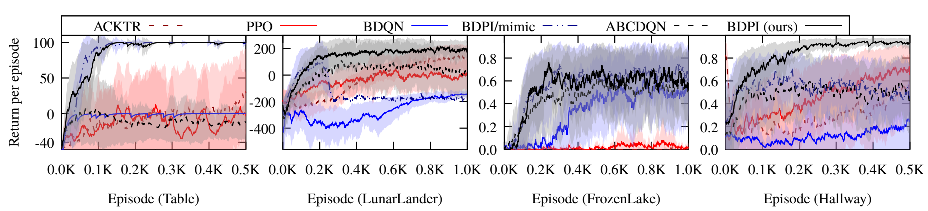

Fig. 1. BDPI outperforms many other algorithms in hard-to-explore, highly-stochastic and pixel-based environments.

$$V(s') \equiv \min_{l=A,B} Q^{l,i}\big(s', \text{argmax}_{a'} Q^{A,i}(s', a')\big)$$
$$\pi(s) \leftarrow (1-\lambda)\pi(s) + \lambda\Gamma(Q^{A,i}(s,\cdot)) \qquad \forall\, i, \forall\, s \in B_i \quad (2)$$

with $Q^{A,i}$ and $Q^{B,i}$ the two Clipped DQN Q-functions of critic i, that are swapped every time-step, and Γ the greedy function, that returns the action having the largest Q-Value in a given state.

3 Experiment

We compare BDPI to a variety of state-of-the-art reinforcement-learning algorithms in three environments: *Table* [7], *LunarLander* and *FrozenLake* (OpenAI Gym), and *Hallway* [1]. Figure 1 shows that BDPI largely outperforms every other algorithm, even in the pixel-based 3D *Hallway* environment. More importantly, BDPI outperforms *ABCDQN*, the critics of BDPI used with no actor, and *BDPI/mimic*, that uses a different actor training rule [7]. This demonstrates that both our actor and critic learning rules advance the state of the art in sample-efficient reinforcement learning. Our results are further illustrated by our robotic wheelchair demonstration, also submitted to this conference.

References

1. Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17:39:1–39:40, 2016.
2. Matteo Hessel, Joseph Modayil, Hado van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Daniel Horgan, Bilal Piot, Mohammad Gheshlaghi Azar, and David Silver. Rainbow: Combining improvements in deep reinforcement learning. *Arxiv*, abs/1710.02298, 2017.
3. Vijaymohan R. Konda and Vivek S. Borkar. Actor-Critic-Type Learning Algorithms for Markov Decision Processes. *SIAM Journal on Control and Optimization*, 38(1):94–123, jan 1999.
4. Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv*, abs/1801.01290, 2018.
5. Ziyu Wang, Victor Bapst, Nicolas Heess, Volodymyr Mnih, Rémi Munos, Koray Kavukcuoglu, and Nando de Freitas. Sample Efficient Actor-Critic with Experience Replay. Technical report, 2016.
6. Bruno Scherrer. Approximate policy iteration schemes: A comparison. In *Proceedings of the 31th International Conference on Machine Learning (ICML)*, pages 1314–1322, 2014.
7. 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.
8. Scott Fujimoto, Herke Van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International Conference on Machine Learning (ICML)*, pages 1582–1591, 2018.
9. Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped DQN. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[1] `https://github.com/maximecb/gym-miniworld`