CURIOSITY AUGMENTED EXPLORATION FOR OFF-POLICY REINFORCEMENT LEARNING

Ashish Malik
Department of Mechanical Engineering
Punjab Engineering College
Chandigarh, 160012, India
ashishmalik.bemech14@pec.edu.in

ABSTRACT

Boltzman exploration is a standard tool in reinforcement learning and is actively used in state of the art learning algorithms. These algorithms still suffer from poor sample efficiency which prohibits direct learning on real hardware. We tackle this problem by addressing the already identified short-comings of Boltzman exploration and shortcommings of recent alfo using a novel method that drives exploration using a combination of intrinsic and extrinsic rewards. We show that our method achieves state of the art sample efficiency in popular high-dimensional continuous control benchmarks.

CHANGE EVERYTHING. PESSIMISTIC UNDEREXPLORATION IS CAUSED BY ALGO, NOT BOLTZMANN.

1 INTRODUCTION

State of the art reinforcement learning algorithms such as Soft Actor Critic (SAC, Haarnoja et al. (2018)), Twin Delayed Deep Deterministic policy gradient (TD3, Fujimoto et al. (2018)), Proximal Policy Optimization (PPO, (Schulman et al., 2017)), etc. are actor-critic methods that employ Boltzmann exploration for generating new trajectories for learning. These algorithms achieve great success in challenging high-dimensional continuous control problems. The off-policy learning algorithm, SAC provides the greatest sample efficiency among these algorithms. It maximizing a trade-off between expected extrinsic returns and entropy of the learned soft policy. However, SAC still requires millions of environment interactions for learning viable policies and value function estimates. Poor sample-efficiency is a major obstacle for widespread adoptation of deep reinforcement learning (DRL) for real world tasks. Two major shortcomings have been identified (Ciosek et al., 2019) in current DRL methods which contribute to their poor sample-efficiency.

1. *Directionally uninformed exploration*: The standard for sampling actions for exploration in DRL is sampling from a Gaussian distributions of action probabilities. The approach is very effective for stable learning. But this sampling results in actions from the opposite side of the mean to be sampled with equal probabilities. These high frequency consequentive perturbations acts as a low pass filter and tend to cancel each other, leading to poor exploration (Stulp & Sigaud, 2013; Kober & Peters, 2008). Also, current policies are obtained after incremental updates over past policies. Therefore directionally uniform exploration is wastefull as actions spaces where past policies had high probability densities have likely been explored.
2. *Pessimistic underexploration*: SAC and TD3 stabilize learning by greedy maximization of a lower bound of action-value estimates to avoiding overestimation (Van Hasselt et al., 2015). However, a spurious maximum of this lower bound will reduce policies covariance at those state-action values. This discourages the policy for exploring action spaces which may improve the critics estimation.

In this work, we propose a novel method for improving the sample-efficiency of off-policy learning algorithms. Our method disentangles the exploration and exploitation trade-off in off-policy learning algorithms and we name it NAME. NAME uses a separate exploration policy which augments Boltzmann exploration with intrinsic rewards. We explore how this augmentation mitigates

undirected explorations and *pessimistic underexploration* (Ciosek et al., 2019) by using only an additional hyperparameter that explicitly controls the exploration-exploitation trade-off.

We test NAME on several challenging continuous control benchmarks and show that it achieves state of the art sample efficiency. Empirical results demonstrate that NAME improves exploration, rate of policy learning and often leads to better final performance compared to vanilla SAC. We also perform ablations to isolate the effect of the new hyperparameter introduced and demonstrate the stability of NAME.

2 PRELIMINARIES

The standard for sampling actions for exploration in DRL is sampling from a gaussian distributions of action probabilities. The approach is very effective for stable learning.

Thus, consecutive perturbations act as a low pass filter and may cancel each other, leading to poor exploration. (This can be addressed using shifted mean) (Stulp & Sigaud, 2013; Kober & Peters, 2008)

REFERENCES

Kamil Ciosek, Quan Vuong, Robert Loftin, and Katja Hofmann. Better exploration with optimistic actor critic. In *Advances in Neural Information Processing Systems*, pp. 1787–1798, 2019.

Scott Fujimoto, Herke Van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. *arXiv preprint arXiv:1802.09477*, 2018.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*, 2018.

Jens Kober and Jan Peters. Policy search for motor primitives in robotics. *Advances in neural information processing systems*, 21:849–856, 2008.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Freek Stulp and Olivier Sigaud. Robot skill learning: From reinforcement learning to evolution strategies. *Paladyn, Journal of Behavioral Robotics*, 4(1):49–61, 2013.

Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. *arXiv preprint arXiv:1509.06461*, 2015.