Safe Exploration in Deep Reinforcement Learning with Action Priors

Sicelukwanda Zwane[1,2], **Tlou Boloka**[1,2], **Ndivhuwo Makondo**[1], **Benjamin Rosman**[1,2]
University of the Witwatersrand[1], Council for Scientific and Industrial Research (CSIR)[2]
South Africa
{szwane2,tboloka,nmakondo,brosman}@csir.co.za

Abstract

Behaviour learning in deep reinforcement learning is inherently unsafe because untrained agents typically have to sample actions from randomly initialized task policies and from random exploration policies. executing these actions in physical environments can lead agents to harmful states, possibly causing damage and poor initial performance. In this work, we address this problem by using transfer learning to develop a framework for safe reinforcement learning in continuous environments. We show that our exploration policy results in fewer collisions with the environment, better initial performance and earlier convergence compared to the vanilla ϵ-greedy random exploration policy.

1 Introduction

Humans have prior information about the world that helps constrain exploration when learning new tasks. For example, humans are able to learn to play unseen video games faster when using prior expert knowledge such as 'ladders can be climbed, doors can be opened', etc. [Dubey et al., 2018]. This prior knowledge not only reduces the search space of actions to consider, but it also lowers the chances of the human executing actions that may lead to harmful states. In a similar way, we attempt to extract useful common knowledge from expert agents to construct a safer exploration policy for agents solving similar tasks in the same environment. Transferring this prior knowledge amounts to transferring behavioural advice from experts such as how to avoid obstacles, resulting in a safer reinforcement learning agent.

2 Methodology

In this work we consider finite episodic tasks described by the continuous Markov decision process $M = \langle D, R \rangle$ given by the reward function R and the domain $D = \langle S, A, T, \gamma \rangle$, where $S \subset \mathbb{R}^M$ is the M-dimensional state space, $A \subset \mathbb{R}^N$ is the N-dimensional action space, T is the transition function, and γ is the discount factor.

We represent the safe exploration policy using the action priors framework [Rosman and Ramamoorthy, 2012], a distribution over near-optimal trajectories sampled from expert agents. Agents learning new tasks in the same domain would then sample from this distribution during exploration in an ϵ-greedy fashion. In this work, we approximate our action prior distribution over experts using Gaussian processes. The use of Gaussian processes allows us to extend this framework to continuous control domains. We train an expert agent for each task and execute a rollout on each expert task policy to build a dataset of trajectories $\mathcal{D} = \{\tau_i\}_{i=1}^{N}$, where $\tau = \{(s_0, a_0), (s_1, a_1), ..., (s_T, a_T)\}$ and fit a Gaussian process over the data to obtain an exploration policy $\mathcal{P}(a|s)$ (see figure 1).

Submitted to the Black in AI workshop at the 32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

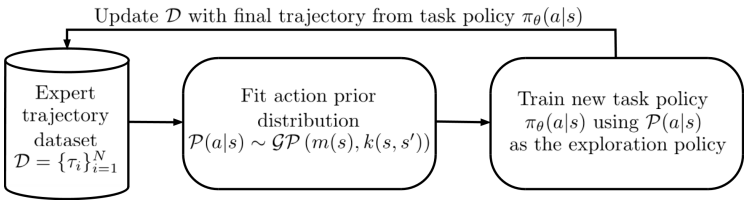
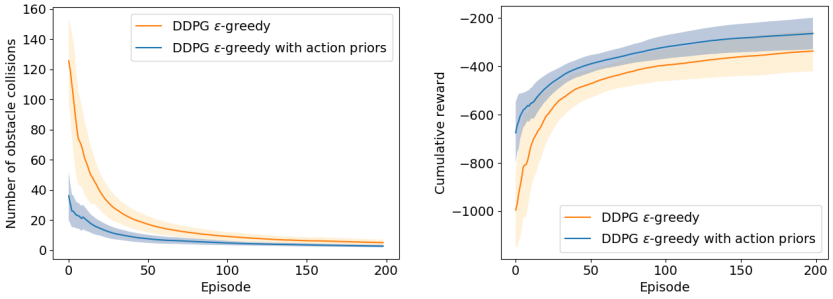


Figure 1: Gaussian process action prior framework



(a) Number of collisions per episode (b) Cumulative reward per episode

Figure 2: Experimental results averaged over 10 random seeds.

3 Experiments

We evaluate our framework by learning policies for navigating around obstacles in a continuous 2D gridworld environment, where states are (x, y) positions, actions are bounded steps (i.e. $[-1, 1]$) along each dimension, and obstacles and walls are regions of high negative reward. Different tasks in this environment have different start and goal states. To learn each task, we use the DDPG algorithm [Lillicrap et al., 2015] with parameter space exploration [Plappert et al., 2017]. [1] We train a new DDPG agent with and without the action prior exploration policy on a test task to evaluate transfer and safety. In this environment, the safety criteria is the number of obstacle collisions per episode.

Preliminary results from the experiments show that there is some benefit to using the action prior exploration policy. In our test tasks, we observe an increase in performance as well as a jump start [Taylor and Stone, 2009] in performance, implying that our modified agent has a better overall task policy(see figure 2b). This shows that our action-prior policy is a viable technique for transfer in continuous control environments. We also observe a decrease in the number of collisions with the environment (see figure 2a), this shows that exploring using our action prior policy results in a safer reinforcement learning agent.

4 Conclusion and Future work

In deep reinforcement learning algorithms that rely on experience replay, the quality of the task policy depends on the quality of samples discovered by the exploration policy. This work combines data from previous expert policies in the same domain to obtain an exploration policy that performs better than a random exploration policy. The benefits of using our action prior policy include safe exploration, a jump start in performance when learning, and increased performance.

Learning a state-based action prior distribution implies that the exploration policy is only useful around previously visited states. This limitation is worse for tasks defined in continuous environments since it is nearly impossible for all experts to have collectively seen all states. In future work, we propose to overcome this restriction using perception-based action priors [Rosman and Ramamoorthy, 2012], where the action prior distribution is conditioned on observations instead of states, since there may be a finite set of observational patterns even though the state space is infinite [Rosman and Ramamoorthy, 2012].

[1] We use the OpenAI baselines implementation by Dhariwal et al. [2017].

References

P. Dhariwal, C. Hesse, O. Klimov, A. Nichol, M. Plappert, A. Radford, J. Schulman, S. Sidor, and Y. Wu. Openai baselines. `https://github.com/openai/baselines`, 2017.

R. Dubey, P. Agrawal, D. Pathak, T. L. Griffiths, and A. A. Efros. Investigating human priors for playing video games. *arXiv preprint arXiv:1802.10217*, 2018.

T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

M. Plappert, R. Houthooft, P. Dhariwal, S. Sidor, R. Y. Chen, X. Chen, T. Asfour, P. Abbeel, and M. Andrychowicz. Parameter space noise for exploration. *arXiv preprint arXiv:1706.01905*, 2017.

B. Rosman and S. Ramamoorthy. What good are actions? accelerating learning using learned action priors. In *Development and Learning and Epigenetic Robotics (ICDL), 2012 IEEE International Conference on*, pages 1–6. IEEE, 2012.

M. E. Taylor and P. Stone. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10(Jul):1633–1685, 2009.