Criticality-Based Advice in Reinforcement Learning (Student Abstract)

Yitzhak Spielberg and Amos Azaria

Computer Science, Ariel University, Israel
yspielb@gmail.com, amos.azaria@ariel.ac.il

Abstract

One of the ways to make reinforcement learning (RL) more efficient is by utilizing human advice. Since human advice is expensive, the central question in advice-based reinforcement learning is, how to decide in which states the agent should ask for advice. To approach this challenge, various advice strategies have been proposed. Although all of these strategies distribute advice more efficiently than naive strategies, they rely solely on the agent's estimate of the action-value function, and therefore, are rather inefficient when this estimate is not accurate, in particular, in the early stages of the learning process. To address this weakness, we present an approach to advice-based RL, in which the human's role is not limited to giving advice in chosen states, but also includes hinting a-priori, before the learning procedure, in which sub-domains of the state space the agent might require more advice. For this purpose we use the concept of critical: states in which choosing the proper action is more important than in other states.

Introduction

The learning process of Reinforcement Learning (RL) agents in complex environments is often very slow. One extensively utilized way to speed up this process is by providing advice to the learning agent by a human teacher. A major downside of this approach is that human advice is expensive. In practice this means that the available advice budget is very limited and thus, one of the the central challenges of advice-based RL is to find a strategy for the efficient selection of advice states (states in which the agent asks for advice) .

In most advice strategies found in the literature, the criteria used for selecting advice states are based solely on *the agent's* model of the policy or the Q-function. In uncertainty-based advice (Da Silva et al. 2020), for example, the selection criterion is the variance of the head outputs of the multi-headed Q-function model. Although advice strategies that use this type of criteria are usually more efficient than primitive advice strategies, such as distributing advice randomly or asking for advice in every state until the advice budget is finished, all of these strategies suffer from a major problem: they are based only on the current understanding of the task *by the agent* (which is represented in the agent's Q-function or its policy). This is a crucial fact because the agent's understanding of the task can be rather poor—especially during the early stage of the learning process. Consequentially, it is likely that in the early stages of the learning process the these strategies will be rather poor at selecting those states in which advice would be most helpful.

The approach proposed in this paper addresses the weakness of most advice strategies mentioned above by including the human expert into the advice framework more extensively. Whereas, in most advice strategies the expert is utilized solely for giving action advice in individual states, in the suggested approach the expert has the additional role to mark sub-domains of the state space in which there might be a strong need for advice. That is, the learning agent utilizes the human expert in two ways: Firstly, to receive advice in individual states; Secondly, to help selecting states in which to ask for advice.

In order to determine states in which advice might be very helpful, we use the concept of state criticality that was introduced in (Spielberg and Azaria 2019). State criticality is a subjective measure of variability in the expected return with respect to the the available actions. The general rule that dictates how to assign a criticality level to a given state is, that states that have a high variability in the expected returns should have a high criticality level while states with low variability in the expected returns should have a low criticality level.

Criticality-Based Advice

While expert advice helps RL agents to learn more efficiently, it is also rather expensive. Hence, there is a need for strategies that select states in which advice is most useful. There exists a variety of techniques that are used to execute this selection task. However, most of them only utilize the agent's knowledge and are therefore rather inefficient - particularly in the early stages of the learning process. The approach that we propose, in contrast, also uses state criticality, which is an aspect of a human's knowledge about the learning environment.

Criticality-based advice is, strictly speaking, not an advice strategy, but a meta-strategy that can be put on top of any underlying advice strategy to make it more efficient. Criticality-based advice utilizes a criticality function

(a function that assigns a criticality level to every state in the environment) that is designed by a human expert a-priori—before the beginning of the RL agent's learning process. In criticality-based advice, advice states are selected by using the criticality function in combination with the selection criterion of the underlying advice strategy. More specifically, a state that has been selected for advice by the underlying advice strategy will receive advice only if it has a sufficiently high criticality level. Thereby, the agent avoids wasting valuable advice on states in which the choice of the proper action has an insignificant influence on the total reward.

The fact that criticality-based advice uses a combination of the metric of the underlying advice strategy — the metric that is used for the state selection criterion of the underlying advice strategy — and state criticality to select advice states leads to the question about the appropriate way how to perform such a combination. One way was suggested above: using the logical *and* operator (*logicand* approach). An interesting alternative way could be multiplication: to multiply the metric of the underlying advice strategy with state criticality. For this type of combination, the selection thresholds for agent uncertainty and state criticality should be fused into one threshold by multiplication too. For this paper, both approaches—the *logicand* approach and the multiplicative approach—were tested (see "Experiments" section).

Experiments

We performed a series of experiments to prove the efficiency of criticality-based advice. First we want to mention the major settings of the experiments :

- All experiments were performed in the Atari Pong environment
- The criticality function was the same as in (Spielberg and Azaria 2019)
- The underlying advice strategy that was used in the experiments presented in this paper is uncertainty-based advice (Da Silva et al. 2020) with an uncertainty treshold of 0.04.
- To enable a fair comparison all advice strategies operated with the same advice budget of 150K.

We tested the 2 versions of criticality-based advice that were mentioned previously: the *logicand* version (BDQN-crit1) and the multiplicative version (BDQN-crit2).Moreover, to evaluate the efficiency of criticality-based advice, two baseline strategies were used. The first strategy was BDQN without advice (BDQN-plain), and the second was BDQN with uncertainty-based advice (BDQN-adv). Both strategies were tested experimentally.

To compare the learning curves of the different advice strategies, every strategy was executed 5 times—each time with a different random seed — and the learning curves of the individual runs were synthesized into one single learning curve by averaging.

The learning curves of the various strategies can be seen in fig. 1.There are several notable observations that can be made upon a closer look at the plot. Firstly, the plot shows that BDQN-adv outperformed BDQN-plain. This anticipated result confirms the usefulness of advice in the Atari Pong environment. The second observation is related to BDQN-adv and BDQN-crit1. It can be seen from the plot, that BDQN-crit1 beats BQQN-adv in the early stages of the learning process but performs slightly worse than BDQN-adv in the later stages. The third remarkable observation is that BDQN-crit2 strongly outperformed BDQN-crit1.

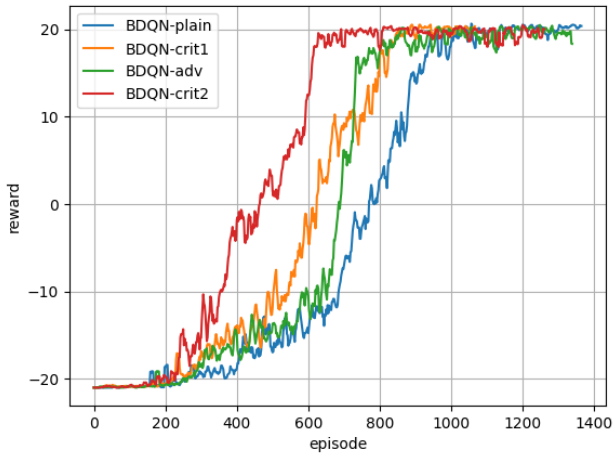

Figure 1: Learning curves for the various advice strategies.

Discussion & Conclusion

The main conclusion that can be derived from the conducted experiments is that augmenting advised-based RL strategies with criticality-based selection criteria is an efficient way to speed up the agent's learning process. We tested two variants of criticality-based advice: the *logicand* variant (BDQN-crit1) and the multiplicative variant (BDQN-crit2). Both variants outperformed the underlying advice strategy, which did not use state criticality. Although the current paper investigated criticality-based advice only for one particular underlying advice strategy and in one particular learning environment, the results indicate that, in general, criticality-based advice is a promising method for speeding up learning for advice-based RL algorithms.

Acknowledgements

This work is supported, in part, by the ministry of science, technology and space, Israel.

References

Da Silva, F. L.; Hernandez-Leal, P.; Kartal, B.; and Taylor, M. E. 2020. Uncertainty-Aware Action Advising for Deep Reinforcement Learning Agents. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04): 5792–5799.

Spielberg, Y.; and Azaria, A. 2019. The Concept of Criticality in Reinforcement Learning. *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, 251–258.