Myopic Behavior in Goal-reaching Reinforcement Learning

Project Proposal for Master Thesis

Motivation

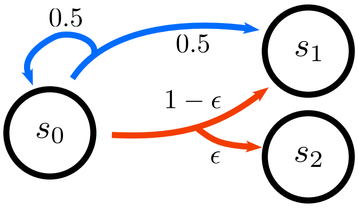

Figure 1: Let us consider the task of reaching s_1 from s_0 in this MDP. The blue action is equally likely to succeed and to loop on the starting state. The red action has a high chance of reaching s_1, with a small risk of reaching the sink state s_2. By repeating the blue action, the probably to succeed tends to 1 as the number of repeats increases. However, under any discounting factor $\gamma < 1$, there exist a value of ϵ such that the optimal policy in the MDP would take the red action, which may trap the agent in the sink state indefinitely.

Goal-reaching problems are a special case within reinforcement learning, in which the reward function is commonly defined as an indicator function conditioned on a particular state (the *goal*). In other words, the agent is only rewarded when a particular state of the environment is reached. Despite the loss of expressivity, this remains a very general framework with wide application [Schaul et al., 2015, Andrychowicz et al., 2017, Ibarz et al., 2021]. However, when coupled with the conventional discounted reward scheme, this reward formulation induces an interesting phenomenon, which can be described as *myopia*: optimization of the RL objective forces the agent to prefer solutions that attempt to reach the commanded goal faster, while risking not to achieve the goal at all. While this phenomenon has been studied to some extent in the more general context of logic-conditioned RL, a better understanding would also greatly benefit goal-reaching algorithms. Notably, in logic-conditioned RL, eventual discounting has been proposed as a solution [Voloshin et al., 2023], but it introduces suboptimality and may be largely impractical.

Scope of the Project

The main goal of this project is to formally and clearly identify properties of MDPs that induce myopic behavior, starting from tabular settings and possibly scaling to continuous ones. Ideally, this study will produce an adaptation scheme for the discount factor, ensuring retrieval of fast policies with optimal success rates. While this project will mostly involve a formal study of the problem, programming skills are necessary for preparing and running numerical simulations.

Tasks

- reviewing related literature
- formally analyzing the emergence of myopic policies
- validating the analysis empirically

Contact

Please, contact Marco Bagatella (mbagatella@ethz.ch) with your CV and a short description of why you find this project interesting.

References

[Andrychowicz et al., 2017] Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., McGrew, B., Tobin, J., Pieter Abbeel, O., and Zaremba, W. (2017). Hindsight experience replay. *Advances in Neural Information Processing Systems.*

[Ibarz et al., 2021] Ibarz, J., Tan, J., Finn, C., Kalakrishnan, M., Pastor, P., and Levine, S. (2021). How to train your robot with deep reinforcement learning: lessons we have learned. *The International Journal of Robotics Research*, 40(4-5):698–721.

[Schaul et al., 2015] Schaul, T., Horgan, D., Gregor, K., and Silver, D. (2015). Universal value function approximators. In *International Conference on Machine Learning.*

[Voloshin et al., 2023] Voloshin, C., Verma, A., and Yue, Y. (2023). Eventual discounting temporal logic counterfactual experience replay. In *International Conference on Machine Learning.*