To the Max: Reinventing Reward in Reinforcement Learning

Grigorii Veviurko, Wendelin Böhmer, and Mathijs de Weerdt

Delft University of Technology

Abstract. This is a "Type B: Encore abstracts" submission based on an article published in Proceedings of the 41st International Conference on Machine Learning (ICML2024) `https://proceedings.mlr.press/v235/veviurko24a.html`.

1 Introduction

Reinforcement learning (RL) is a powerful framework for training agents to make decisions in complex environments through trial and error. The central objective in RL is to maximize the cumulative return, which is a discounted sum of rewards an agent receives while interacting with its environment. However, designing effective reward functions remains a significant challenge. Different reward functions can lead to vastly different learning outcomes, even if they all correspond to the same optimal policy [1]. Some reward functions may drive the agent to efficiently solve the task, while others can result in suboptimal behavior or even cause the agent to get stuck in local optima.

The challenge of reward design is particularly relevant for environments with *sparse rewards*, such as goal-reaching tasks, where the agent only receives rewards upon successfully reaching the goal. As these environments usually can not be solved with traditional methods, a popular approach is to introduce a surrogate dense reward that guides the agent towards the goal [2]. This approach is very sensitive and should be carefully tuned on the specific problem at hand.

To address the challenge of learning in goal-reaching environments, we propose an alternative RL paradigm —- max-reward RL -— where the agent is trained to maximize the highest reward achieved during an episode rather than the cumulative reward. Compared to standard RL, this approach has several appealing theoretical properties: it makes the reward design more intuitive; it is more prone to getting stuck in local optima; and it utilizes reward efficiently due to different bootstrapping scheme. For experimental evaluation, we implement PPO[3] and TD3[4] for the max-reward case and demonstrate their dominance over the cumulative RL counterparts using several goal-reaching environments.

2 Our contributions

To make the max-reward RL approach viable, an analog of the Bellman equation is required, as it is a key component of all RL methods. Prior attempts to derive

such an equation for the maximum reward have either been mathematically incorrect [5] or limited to fully deterministic cases [6]. As a result, the idea of max-reward RL has gone largely unnoticed by the broader community. In our work, we pursue three main goals: a) provide an intuition for why max-reward RL can be a better choice for certain problems; b) establish a theoretical framework for max-reward RL; and c) experimentally demonstrate the benefits of using max-reward RL.

2.1 Intuition: different bootstrapping

We begin by developing an intuition for why max-reward RL can be beneficial. We run a simple tabular experiment on a chain environment, where the agent obtains a high reward for reaching the goal state and a moderate reward for getting halfway to the goal. We run the max-reward and standard value iteration algorithms and compare how fast they learn the optimal policy.

The results demonstrate the the max-reward RL is more efficient, especially when the intermediate reward is smaller. The explanation for that is the difference in bootstrapping: in standard RL, the target for the $q-$value is a sum of the immediate reward and the $q-$value at the next timestep. Therefore, this target changes in each epoch until convergence. In the max-reward case, on the other hand, the target in the max-reward state is just the reward and does not change with time. This example suggests that the max-reward approach is a better choice in environments where the task of the agent is to reach the goal state.

2.2 Mathematical framework

We provide a theoretically sound formulation of max-reward RL. Specifically, we obtain the following results:

- We define max-reward value functions using an auxiliary variable that tracks maximum-so-far reward. These functions satisfy Bellman-like equations.
- We show that the max-reward value functions are unique fixed points of the corresponding Bellman operators and hence can be learned.
- We prove on- and off- policy gradient theorems, thereby enabling various actor-critic methods to be used for max-reward RL.

2.3 Experiments

We implement and evaluate max-reward and cumulative versions of TD3 and PPO algorithms. Using Maze and Fetch environments [2], we demonstrate the benefits of max-reward RL. In the Maze environment, the max-reward algorithms can learn the policy for different "imperfect" surrogate reward functions, while standard methods require a better tuned surrogate. In Fetch domain, max-reward TD3 solves the *Push* and *Slide* tasks, while standard TD3 can not learn anything at all.

References

1. Andrew Y. Ng, Daishi Harada, and Stuart J. Russell: Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping. Proceedings of the Sixteenth International Conference on Machine Learning (1999).
2. Rodrigo de Lazcano and Kallinteris Andreas and Jun Jet Tai and Seungjae Ryan Lee and Jordan Terry, 2023. Gymnasium Robotics. `http://github.com/Farama-Foundation/Gymnasium-Robotics`
3. John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, Oleg Klimov (2017). Proximal Policy Optimization Algorithms. CoRR abs/1707.06347
4. Fujimoto, S., Hoof, H.V., & Meger, D. (2018). Addressing Function Approximation Error in Actor-Critic Methods. International Conference on Machine Learning.
5. Quah, K.H., & Quek, H.C. (2006). Maximum reward reinforcement learning: A non-cumulative reward criterion. Expert Syst. Appl., 31, 351-359.
6. Gottipati, S.K., Pathak, Y., Nuttall, R., Sahir, Chunduru, R., Touati, A., Subramanian, S.G., Taylor, M.E., & Chandar, S. (2020). Maximum Reward Formulation In Reinforcement Learning. ArXiv, abs/2010.03744.