Convergence of Average-Reward Reinforcement Learning

Kasper Engelen
kasper.engelen@uantwerpen.be

Guillermo A. Pérez

Academic year 2024 – 2025

1 Introduction

Reinforcement learning is a machine learning technique for sequential decision making in unknown and stochastic environments. The learning process consists of taking actions and observing a reward. The result is an optimal policy that tells the system what actions are optimal in which situations, taking into account both immediate and long-term rewards. Such an optimal policy maximises a function of the accumulated rewards.

Two such functions that are commonly considered are discounted-reward and average reward. Much is known about the discounted-reward setting, resulting in popular algorithms such as Q-learning. The average-reward setting remains less well explored, with many algorithms lacking convergence guarantees. Despite being less popular, the average-reward setting has the advantage that you do not need to specify a discounting factor, which is often difficult to estimate.

In this project you will explore the literature on average-reward reinforcement learning algorithms. More specifically you will read and reproduce the results from a paper by Wan et al. [4], which introduced an average-reward reinforcement learning algorithm with convergence guarantees called Differential Q-learning.

Prior theoretical knowledge on reinforcement learning and programming experience are both strongly recommended. During the project you will obtain theoretical insight into reinforcement learning, practical implementation skills, as well as skills related to writing and presenting research. As such, the project forms an interesting topic for a master's thesis.

2 Project outline

You will first have to familiarise yourself with the theoretical foundations of reinforcement learning, and in particular the foundations of average-reward reinforcement learning [1, 2]. Then, you will reproduce the research of the paper by Wan et al. titled "Learning and Planning in Average-Reward Markov Decision Processes", by (a) exploring the problem setting, including related and prior work, (b) reproducing the theoretical results (i.e., the theorems and proofs), and (c) implementing the algorithms from the paper.

Throughout the project, you will showcase your findings by giving presentations to the research

group. Finding and presenting a link between the paper by Wan et al. on one hand, and R-learning [3] and stochastic approximation [5] on the other hand, would be particularly interesting.

References

[1] B. V. Houdt and G. A. Pérez. *Mathematical Foundations of Reinforcement Learning.*

[2] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. 1st. USA: John Wiley & Sons, Inc., 1994. ISBN: 0471619779.

[3] A. Schwartz. "A Reinforcement Learning Method for Maximizing Undiscounted Rewards". In: *Machine Learning, Proceedings of the Tenth International Conference, University of Massachusetts, Amherst, MA, USA, June 27-29, 1993*. Ed. by P. E. Utgoff. Morgan Kaufmann, 1993, pp. 298–305. DOI: 10.1016/B978-1-55860-307-3.50045-9. URL: https://doi.org/10.1016/b978-1-55860-307-3.50045-9.

[4] Y. Wan et al. "Learning and Planning in Average-Reward Markov Decision Processes". In: *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*. Ed. by M. Meila and T. Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 10653–10662. URL: http://proceedings.mlr.press/v139/wan21a.html.

[5] Wikipedia contributors. *Stochastic approximation — Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/w/index.php?title=Stochastic_approximation&oldid=1189125221. [Online; accessed 15-March-2024]. 2023.