# Formal Foundations of Model-Free 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.

Q-learning is an algorithm for reinforcement learning, that can be used to obtain an optimal policy that maximises the so-called expected infinite-horizon discounted reward [2, 3]. Given that reinforcement learning is also being applied in safety-critical contexts, such as energy-management and healthcare, it is essential that the Q-learning algorithm is provably correct.

Originally, the (asymptotic) correctness of the Q-learning algorithm follows from the correctness proof of the Robbins-Monro scheme for stochastic approximation, which is the technique that underlies Q-learning [4]. Nowadays, however, more modern treatments of stochastic approximation exist, in particular those that leverage ODE theory [1]. In this project you will research such modern treatments (e.g., course texts, proofs, and mathematical literature) and incorporate them into the correctness proof of Q-learning.

Relevant skills for this project are knowledge of ODEs and stochastics, as well as general skills regarding scientific research such as writing, presenting, and studying literature.

## 2 Definitions

**Q-learning.** This algorithm will learn a so-called Q-function that gives the expected infinite-horizon discounted reward, for a specific state $s$ and action $a$, assuming that the optimal policy $\pi^*$ is followed from then onwards. The Q-values are defined as

$$Q^*(s,a) = \mathbb{E}_{\pi^*}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s, a_0 = a\right].$$

These values can be obtained by iteratively updating the Q-values

$$Q'(s,a) = (1-\alpha) \cdot Q(s,a) + \alpha \left( r + \gamma \cdot \max_{a'} Q(s',a') \right),$$

where $a$ is the action taken in state $s$, $r$ is the observed reward, and $s'$ is the observed resulting state. The parameters $\gamma$ and $\alpha$ are the discounting factor and learning rate, respectively. $Q(s,a)$ will then asymptotically converge to $Q^*(s,a)$, for all $s$ and $a$.

**Stochastic approximation.** Stochastic approximation methods are iterative methods for root-finding and optimisation in stochastic settings. We wish to find the root of a function $f(\theta) = \mathbb{E}_X[F(\theta, X)]$, that depends on the random variable $X$. This function cannot be evaluated directly and can only be observed through stochastic measurements. The goal is then to find a root $\theta^*$ such that $f(\theta^*) = m$ with $m \in \mathbb{R}$.

**Robbins-Monro algorithm.** One specific method to solve the stochastic approximation problem is the Robbins-Monro scheme. The Robbins-Monro scheme is important in machine learning, forming the basis of the widely used Q-learning and stochastic gradient descent algorithms. It is an iterative scheme of the form

$$\theta_{t+1} = \theta_t - \alpha_t y_t.$$

Here, $\{y_t\}_{t=0}^{\infty}$ are samples drawn from a distribution with expectation $E[f(\theta_t, X)]$. The $\{\alpha_t\}_{t=0}^{\infty}$ are step sizes such that $\sum_t \alpha_t = \infty$ and $\sum_t \alpha_t^2 < \infty$. The sequence $\{\theta_t\}_{t=0}^{\infty}$ will converge to $\theta^*$ almost-surely.

# 3 Project Project outline

In the course of this project, you will first read up on reinforcement learning, focusing on the link between the Q-learning algorithm and stochastic approximation. Then, you will more deeply explore modern treatments of stochastic approximation.

In the middle of the project, you will compile your findings into a small modern tutorial on stochastic approximation, and present it to our research group, together with some initial idea on how to apply this technique to the Q-learning algorithm.

Finally, you will incorporate stochastic approximation into the full proof of the Q-learning algorithm and present the full proof.

# References

[1] J. L. Ny. *Dynamic Programming and Stochastic Control.* 2009. Chap. 15.

[2] C. J. C. H. Watkins and P. Dayan. "Technical Note Q-Learning". In: *Mach. Learn.* 8 (1992), pp. 279–292. DOI: `10.1007/BF00992698`. URL: `https://doi.org/10.1007/BF00992698`.

[3] Wikipedia contributors. *Q-learning — Wikipedia, The Free Encyclopedia.* `https://en.wikipedia.org/w/index.php?title=Q-learning&oldid=1193548086`. [Online; accessed 15-March-2024]. 2024.

[4] 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.