Master Project

# On the road from Model-Based Dynamic Programming to Model-Free Reinforcement Learning: a Sample Efficient Approach

Pol Mur i Uribe, MSc Systems & Control, TU Delft
P.MuriUribe@student.tudelft.nl

Pedro Ferreira, Delft Center for Systems and Control, TU Delft
Pedro.Ferreira@tudelft.nl

Peyman Mohajerin Esfahani, Delft Center for Systems and Control, TU Delft
P.MohajerinEsfahani@tudelft.nl

## Context

Reinforcement Learning algorithms can be deployed when a system that needs to be controlled can be posed as a Markov Decision Problem (MDP). MDPs are mathematical models that describe a system where the stochastic state transitions are influenced by a controller. Reinforcement Learning aims to solve the MDP by minimizing its associated cost choosing the optimal action to take, in every possible state of the system. Reinforcement Learning has been a subject of research for the last decades with very successful results in several fields. Some examples of successful applications are TD-Gammon (Backgammon, 1992), AlphaZero (Chess and Go, 2017), and AlphaStar (Starcraft II, 2019).

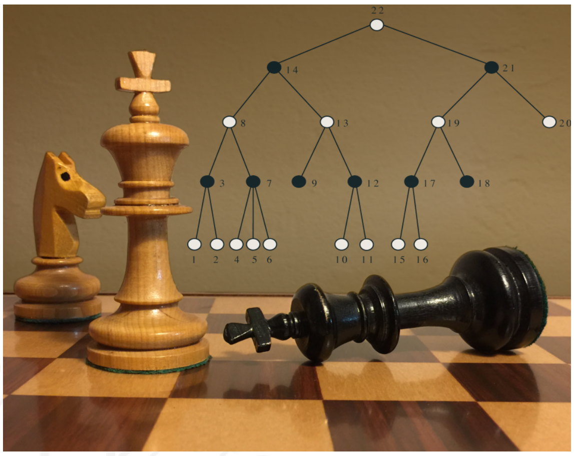

Figure 1: Research Tree for Chess playing RL algorithm (D. Bertsekas, 2021)

Depending on the previous knowledge about the environment we can divide the control approaches between model-free, where the dynamics of the system are unknown, and model-based , where the dynamics are known. Improvements regarding the algorithms proposed in the literature are published every year to try to overcome some of the flaws Reinforcement Learning is facing, such as the high computational and sample complexities. A solution to an MDP is a description of which action a controller chooses given every possible state the system may be in

$$J^*(s) = \min_a \left( c(s,a) + \gamma \sum_{s'} P\left(s,a,s'\right) J^*\left(s'\right) \right),$$

for each state $s$ the solution depends on the stage cost associated to its current state-action pair $c(s,a)$ , a discount factor $\gamma$, the transition probability matrix $P(s,a,s')$ and the predicted expected cost of the future states.

If the MDP (in particular $P(s,a,s')$ and $c(s,a)$) are known, the computation of the cost function can be done iteratively using model-based algorithms such as value iteration or policy iteration. If the MDP is not known, then one has to use model-free techniques that require sampling of the state-action spaces, and this often takes many samples before any model-free algorithm converges close enough. In practice, however, there are instances in which $c(s,a)$ and $P(s,a,s')$ are partially known. To a certain extent, one could disregard the partial knowledge of $c(s,a)$ and $P(s,a,s')$ and use model-free techniques to solve the MDP. This approach, however, removes the possibility of a potential speedup caused by the partial knowledge. By exploiting the prior knowledge about the MDP, a new algorithm could be proposed that would combine the model-free algorithms with the already known values of $c(s,a)$ and $P(s,a,s')$, creating a hybrid between model-based and model-free reinforcement learning. In some applications where the collection of samples is expensive and complex this approach could help reducing the cost of the data collection, a crucial step in model-free approaches.

## Project tasks

This master thesis project is aimed at creating a new reinforcement learning framework in which an MDP is partially known.This new framework would allow for the development of new algorithms that combine model-based and model-free algorithms. The tasks to be completed along the project are:

1. Investigate the fundamentals of Reinforcement Learning and what makes model-based algorithms computationally expensive as well as the importance of sample complexity.

2. Propose an algorithm that could combine the prior knowledge about the MDP with model-free techniques.

3. Investigate the current literature to find the State-of-Art algorithms and compare their performance against that of the proposed algorithm.

4. Find novel applications where the proposed algorithm could be implemented with success.