– Master Project Proposal–

SENSITIVITY ANALYSIS AND INTRINSIC HORIZONS IN MARKOV DECISION PROCESSES

O-A. MAILLARD
SequeL, Inria Lille – Nord de France
E-MAIL: odalric.maillard@inria.fr

Keywords: Reinforcement Learning, Markov Decision Processes, Forecast horizon, Sensitivity analysis.

Introduction Reinforcement Learning is a field of research that models the problem of a learning agent interacting with a partially known dynamical system, that plays action based on past observations and receives reward that the agent tries to maximize. Markov Decision Process is a standard way of modeling a reinforcement learning problem when the observations are states (sufficient statistics of the past observations).

In this project, we consider an MDP with finitely many states and actions, when the learner interact in a single stream of states-actions-rewards with the system, starting from some initial state revealed at the beginning of the game. The total number of interactions T, is the horizon of the learning problem, and can be either finite or infinite. The goal of the learner is to compute a policy (a mapping from states to actions) that ensures it accumulates high enough rewards.

We make two observations: First, the optimal policies in an MDP may differ a lot depending T. Also, an optimal policy for a finite horizon T is typically not-stationary, in the sense that its recommended actions explicitly depend on the number of actions played from the starting point, while an optimal policy for a infinite horizon $T = \infty$ will be stationary in typical MDPs.

Now when considering reinforcement learning for MDPs, the transition distributions and rewards are unknown and estimated from observations. The optimal policies for an infinite horizon are characterized by optimal Bellman equations that make appear two key quantities, the (average) gain and the bias function. It is thus natural to wonder how a noisy estimation of these objects affects the estimation of the gain and of the bias function.

Answering these questions may trigger significant progress in the theoretical and practical understanding of reinforcement learning, and lead to a major progress in the field.

Goal In this project, we want to better understand the effect of the time horizon on the optimal policies, as this dependency has been rarely addressed in the recent literature. An interesting phenomenon is that in many MDPs, provided that T is large enough, the optimal non-stationary policies strategies with finite horizon T will coincide with an optimal policy for infinite-horizon (at least for its initial steps). This gives rise to the notion of Intrinsic Horizon, or Turn-pike of an MDP.

A second goal is to study the sensitivity of the gain and bias function to a perturbation of the transition and reward distributions. This seemingly unrelated question seems indeed strongly connected to the notion of intrinsic horizon and, while the sensitivity of the gain function is well understood, the sensitivity of the bias is more difficult to handle.

Main tasks The goal of this Master project is first to provide a bibliographic overview of the notion of Intrinsic Horizon, Turn-pike and sensitivity analysis in MDPs. Then, to illustrate the effect of the horizon on well-chosen MDPs via numerical experiments. The next step will be to investigate ways to take advantage of the effective horizon, and more importantly to propose ways to estimate it. These ways should be theoretically-grounded as much as possible and supported with key numerical experiments. In parallel, the sensitivity will be investigated in depth, and then combined with the insights provided by the study of intrinsic horizon.

Other information The student should be mathematically strong and interested in solving theoretical problems using probability, statistics and optimization. A prior knowledge of the MDP literature would be a great addition. While the main goal of this internship is to solve a theoretical problem, the student should be able to run some simple numerical experiments to assess the practical performance of the algorithms (in the programming language of his/her choice).

SequeL is an Inria research team based in Lille and specialized in all aspects of sequential decision making, with a rich scientific activity. This research internship proposal is part of a national research project funded by the ANR that focuses on handling non-stationarity and structure in multi-armed bandits.

Bibliographic references Some simple entry points will be given upon request.