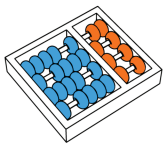
MC935rA/MO436A - Reinforcement Learning
Institute of Computing - Unicamp
Second Semester 2020



Prof. Esther Colombini
esther@ic.unicamp.br

Project 1 - Deadline: 22/12/2020

1 Goal

The goal of this assignment is to apply reinforcement learning control methods to a problem modeled and specified by the group. You must clearly define:

- The problem addressed
- The MDP formulation
- The discretization model adopted

The work consists of finding an adequate solution to the chosen problem, evaluating it according to: computational cost, optimality, influence of reward function, state and action space sizes. You are required to clearly define:

- How the problem was modeled
- Implementation specifics and restrictions

2 Problem

Write an environment that implements your problem. You can adapt environments that are available in the literature, but, in any case, you should fully characterize your environment by defining:

- The nature of your environment (episodic/not episodic, deterministic/stochastic)
- What are your terminal states (when they exist)
- How is your reward function defined
- All parameters employed in your methods (discount factor, step size, etc.)

Your group is required to implement the following methods:

1. **Monte Carlo Control:** Initialize the value function to zero. Use a time-varying scalar step-size of $\alpha_t = 1/N(s_t, a_t)$ and an $\epsilon-$greedy exploration strategy with $\epsilon_t = N0/(N0+N(s_t))$, where $N0$ is a constant, $N(s)$ is the number of times that state s has been visited, and $N(s, a)$ is the number of times that action a has been selected from state s. You should define $N0$ to fit your problem. Plot the optimal value function $V^*(s) = max_a Q^*(s, a)$.
2. **Q-learning (or some variation like DoubleQ-learning):** Initialize the action-value function to zero. Use a random selection of actions whenever you have a draw among actions. Use the same step-size and exploration schedules as MC Control. Define the number of episodes and discount factor accordingly.

3. **SARSA(λ):** Initialize the action-value function to zero. Use a random selection of actions whenever you have a draw among actions. Use the same step-size and exploration schedules as MC Control. Use the same number of episodes as Q-learning. Vary $\lambda \in \{0, 0.2, ..., 1\}$.

4. **Linear function approximator:** propose and apply it to all the above mentioned algorithms.

The system must be evaluated according to the quality of the solutions found and a critical evaluation is expected on the relationship between adopted parameters x solution performance. Graphs and tables representing the evolution of the solutions are expected. Additional comparisons with the literature are welcome, although they are not mandatory.

3 Programming language

You should use Pyhton as programming language. However, interfacing with other languages and libraries is permitted once provided the reasons for such adoption.

4 Evaluation and Discussion

The system should be evaluated according to the quality of the solutions found and a critical evaluation is expected on the relationship between adopted parameters x solution quality. Graphs, tables and images representing the results are expected. Further comparisons with the literature are welcome, although not mandatory. A link to a video of up to 5 minutes with recording of the solution running in the scene should be indicated in the report.

Please, discuss in the report:

- The advantages and disadvantages of bootstrapping in your problem.
- How the reward function influenced the quality of the solution. Was your group able to achieve the expected policy given the reward function defined?
- How function approximation influenced the results. What were the advantages and disadvantages of using it in your problem?

5 Groups

The groups must be composed of 5 members.

6 Report

The definition of the problem, the solution, and the results obtained must be presented in a report created as a Jupyter notebook. Please, make sure you put the graphs, tables, comparisons, and critical analysis in the notebook. The report should clearly indicate what the contribution of each team member was.