TD reinforcement Learning - Univ. Paris-Saclay

Herilalaina Rakotoarison, Laurent Cetinsoy, Diviyan Kalainathan, Michèle Sebag

25 novembre 2021

1 Exercises

Q. 1.1 what is the goal of Reinforcement Learning ? In this framework what is the goal of an agent ? How can you formalize it mathematically ?

Q. 1.2 What make reinforcement learning different from supervised learning ? What makes reinforcement learning hard ?

Q. 1.3 Give the formal definition of a Markov Decision Process (MDP). Give the meaning of it components ?

Q. 1.4 What does the discount factor γ mean or represent ?

Q. 1.5 Can you give exemples of setup with non deterministic transitions ?

Q. 1.6 What is the difference of an episodic setup and a continuous one ? Give an example of each.

Q. 1.7 What does V function means or represents ?

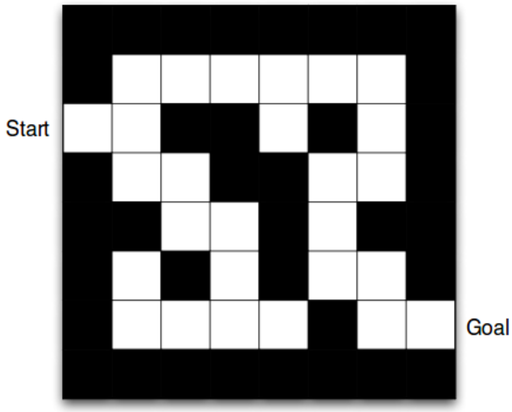
2 Maze / Labyrinth

Let us consider the following labyrinth made of 27 white boxes In order to teach him to find to reach the exit, we give it a reward 0 if is reach the goal and and punition of -1 each time he is on another white square. We consider the environment as deterministic : if the agent decides to go to an adjacent white box, it indeed goes there ; and $\gamma = 0.9$.

Q. *2.1* What is the space of states ?

Q. *2.2* What is the space of actions ?

Q. *2.3* Why do we give a negative reward when the agent does not reach the goal ?



Q. ***2.4*** Give the bellman equation followed by V

Q. ***2.5*** Deduce what is V for each state for the optimal policy (the one where the agent takes the best decision at each step)

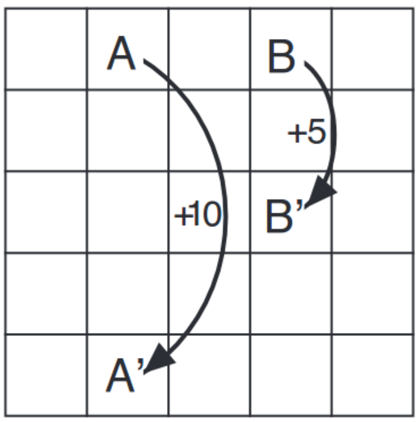
3 Grid

Let us consider the following environment : states : squares(i, j) with $i \in \{1, ...5\}$, $j \in \{1, ...5\}$; actions : go North, South, East, West.

Transitions :

- as expected in general
- if your action brings you outside the state space, you stay where you are (with reward -1)
- if you are in A, any action brings you to A' (with reward 10)
- if you are in B, any action brings you to B' (with reward 5).

Rewards : -1 for going out ; + 10 in A ; + 5 in B ; else 0.



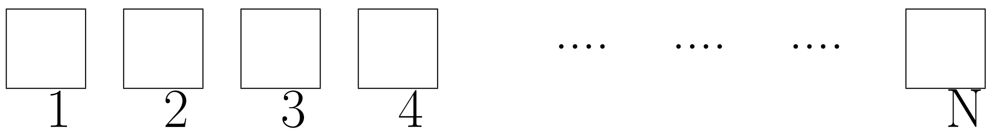
Q. 3.1 We consider the following policy : always going upside. Give the value function for all states. We consider an infinite horizon and we will take $\gamma = 0.9$ as discount factor.

4 River

We want to make an agent cross a river modelised by N boxes. The box number 1 is the starting point and the box N is the goal. The state space is thus 1..N and the action space is left, right the reward function if R(s) = 0 for $s < N$ *and R(N) = 100. State N is terminal. Discount factor* $\gamma = 0.9$

Let the transition function be defined as :

$$\begin{array}{lll} p(s, Right, s+1) = 1 & \textit{if}\ \ s < N & \\ p(s, Left, s-1) = 1 & \textit{if}\ s > 0 & \textit{else}\ p(1, Left, 1) = 1 \end{array}$$

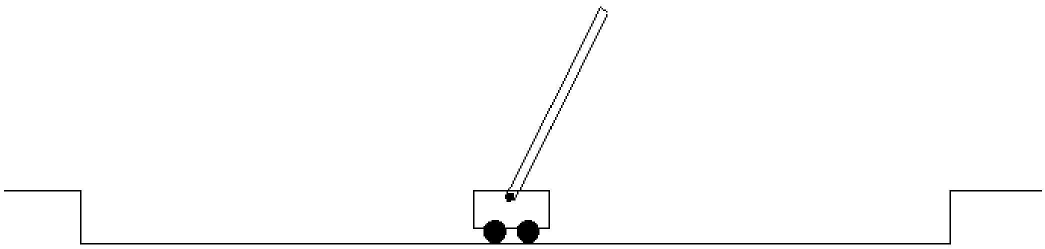


Q. *4.1* Compute the value function (function of N) for a constant policy, $\pi(s) = Right$ for all s.

Q. *4.2* Same question with $p(s, Right, s+1) = .9;\ p(s, Right, s) = .1$.

5 Pole balancing

Let us consider the CartPole problem where on wants to maintain a stick vertically. The stick is fixed at the bottom to a cartpole that can move.



Q. *5.1* Propose a reward structure which would likely induce the desired behavior