Reinforcement Learning – Final exam

Nicolas Gast Panayotis Mertikopoulos

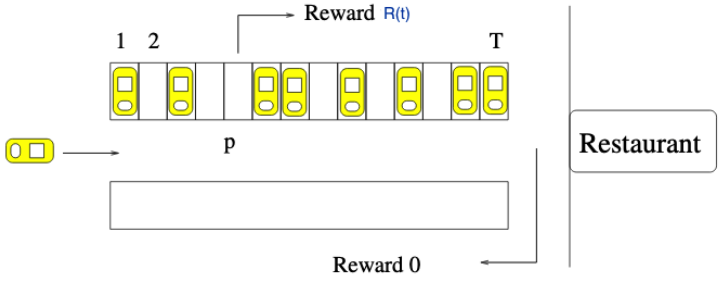
February 2nd, 2022 (academic year 2021-2022)

Duration: 2 hours. You are allowed up to 10 pages of your personal notes. Calculator are authorized (but probably not useful).

Please justify carefully your answer (try to be **concise** and **precise**). The grading scale is given as an indication.

Exercise 1 Markov Decision Processes and Reinforcement Learning (5pts)

A driver wants to park her car as close as possible to the restaurant. There are T parking slots in the street. The driver visits the slots one by one (starting from slot 1 to slot T). Each slot is either available or taken. The driver cannot see if a slot is available unless she is in front of the slot. When she arrives at an empty slot, she can decide to park now or to continue driving and hope to find a slot that is closer to the restaurant. If she chooses slot t, she earns a reward $R(t)$. If she reaches slot T and this slot is not available, she must go home and her reward is 0.



We assume that the probability for each slot to be available is $p \in [0, 1]$.

1. (5pts) We first consider that the driver knows p and wants to maximize her expected reward. For that, she models the problem has a Markov decision process where the state can be any of $(t, \text{available})$ or (r, taken), for $t \in \{1 \dots T\}$. The possible actions are "park" or "continue" for a state $(t, available)$ or "continue" for a state $(t, taken)$. We denote by $\mathcal{S}$ the state space and by $\mathcal{A}(s)$ the set of available actions in state $s \in \mathcal{S}$.
   a) What is the instantaneous reward $r(s, a)$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}(s)$?
   b) What are the probability of transitions $p(s'|s, a)$ for all $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}(s)$?
   c) Consider the policy π^{stop} that parks as soon as an available slot is available, and let $V^{\pi^{stop}}(s)$ be the value function when starting in state s. Find a recursive equation that links $V((t, \text{taken}))$ and $V((t + 1, \text{taken}))$.
   d) Assume that $p = 0.5$, $R(t) = t$ and $T = 10$. Compute $V((t, \text{taken}))$ for all $t \in \{1 \dots T\}$ (if you cannot compute it numerically, indicate an algorithm that can compute it).
   e) By using similar ideas, what is the optimal policy and its value function?

2. (Bonus, +2pts) We now assume that p is unknown to the driver and needs to be learnt while parking. We propose to use a Bayesian approach where the prior distribution of p is uniform. We use a different MDP model where the state is a tuple (t, n, x) where $t \in \{1 \dots T\}$ is the slot, $n \in \{0 \dots t-1\}$ is the number of slots that have been observed empty before slot t and $x \in \{available, taken\}$ indicates if the slot is available or not.
   a) Gives the rewards $r(s, a)$ and probability of transitions $P(s'|s, a)$ of this model.
   b) Derive an algorithm that computes an optimal policy.