EL2805 Reinforcement Learning

Homework 1

November 11, 2025

Division of Decision and Control Systems
School of Electrical Engineering and Computer Science
KTH Royal Institute of Technology

Instructions (read carefully):

- Solve Problems 1 and 2.
- Work in groups of 2 persons.
- **Both** students in the group should upload their scanned report as a .pdf-file to Canvas before November 22, 23:59. The deadline is strict. Please mark your answers directly on this document, and **append** hand-written or typed notes justifying your answers. Reports without justification will not be graded.

Good luck!

1 Modelling and Optimal Control

The objective of this part is to model a practical optimal problem using the Markov Decision Process framework, and to compute (not learn) the optimal control policy.

1.1 Modelling

Model, if this is at all possible, the two following problems using a Markov Decision Process. Precise the time horizon, state and action spaces, the transition probabilities, and the rewards. Do not try to solve these MDPs.

Q1. You get to observe a fair 6-sided die being rolled $T > 21$ times and generating outcomes $X(i)$, $i \in \{1, ..., T\}$. You may stop observing at any time before T and when you do so, denoting by s the sum of all observations up to this point, you receive a reward $s - 21$ if $s \leq 21$ or -10 otherwise.

Q2. A dictator is aiming at sequentially passing a series of N laws $l_1, l_2, ..., l_N$ that would allow him to accumulate wealth. His starting wealth is w_0 and each law will double his wealth, however, each law carries a probability p_r that the people revolt and a probability p_v that the parliament rejects the law. If the parliament rejects a law, the dictator does not gain any wealth (the law cannot be re-introduced). However if the people revolt, the dictator loses all his fortune and is overthrown. The two events (revolt and rejection) are independent. The goal is to retire having accumulated the largest fortune.

1.2 Optimal control

A race car must complete a total of T laps around a circuit. When equipped with new tires, the car can complete a lap in t_0 seconds and loses a random number of seconds $l(n)$ for each lap n in which it uses the same tires (i.e. to complete the m-th lap with the same tires it takes $t_0 + \sum_{i=1}^{m-1} l(i)$ seconds, with $l(i)$ drawn i.i.d. according to a Bernoulli distribution, $\mathbb{E}[l(i)] = 0.5$). At the start of each lap, the driver is allowed to execute a pit stop which takes 25 seconds (added to the lap's time) but allows the car to continue with fresh tires (i.e. it can complete a lap in t_0 seconds again). Design a pit stop strategy that will minimize the expected finishing time of the car.

Q3. *Model this problem as a Markov Decision Process (describe this MDP in detail).*

Q4. *Denote by $P(i)$, the lap when the last pit stop before lap i occurred and denote the wear of tires at lap i by $w_i = \sum_{j=P(i)}^{i-1} l(j)$. Establish that the optimal policy is threshold-based, i.e., at lap i, the optimal action is to make a pit stop if and only if $w_i \geq \alpha_i$ for some threshold α_i.*

Q5. *Compute the optimal policy and expected completion time for the last 2 laps of the race for a wear level of w.*

Q6. *Establish that under the optimal policy, the driver must never perform a pit stop in the last lap.*

2 Delayed Markov Decision Processes

Consider a Markov Decision Process with infinite horizon and discount factor $\lambda \in (0,1)$. It is characterized by its stationary transition probabilities $p(\cdot|s,a)$ and deterministic and bounded reward function $r(s,a)$ for all state-action pair (s,a). The objective is to learn an optimal control policy. Due to communication delays between the agent and the system (or environment), when the agent decides on an action a_t at step t, it then receives, as a feedback, the state s_{t+1-d} and the reward $r_{t-d} = r(s_{t-d}, a_{t-d})$ corresponding to the action she sent to the system d steps ago. We assume that when an action is selected, it is applied without any delay, i.e., s_{t+1} is sampled according to the distribution $p(\cdot|s_t, a_t)$ at any step t. This delayed MDP scenario is depicted in Figure 1. At step t, the agent must decide on a_t based on the information she receives previously $(s_1, a_1, r_1, \ldots, s_{t-d-1}, a_{t-d-1}, r_{t-d-1}, s_{t-d})$. Hence the agent is *blind* before step $d+1$ (she has to select actions without any information about the state).

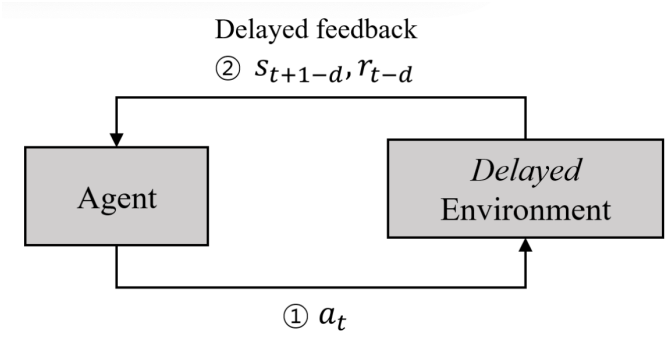

Figure 1: MDP with constant interaction delays. The agent receives delayed information about the state of the system and the rewards.

2.1 Naive markovian policies

Despite the delays, we consider using naive Markovian policies. Under such a policy π, the action selected at step $t \geq d+1$ is a function of s_{t-d} only. At steps $1, \ldots, d$, the agent has no information about the state of the system, and in these steps, she selects action 0. We define by $V^\pi(s)$ the expected discounted reward under policy π when the initial state is s.

Q7. *Try to write fixed point equations satisfied by V^π. Do not try too hard, and explain why this cannot be done.*

2.2 Equivalent augmented state MDP

Consider augmenting the state as follows. Let

$$\bar{s}_t = (s_{t-d}, a_{t-d}, \ldots, a_{t-1}).$$

To have a consistent notation, even when $t \leq d$, we let $(s_{-d}, \ldots, s_0) = (0, \ldots, 0)$ and $(a_{-d}, \ldots, a_0) = (0, \ldots, 0)$ where 0 denotes a 'dummy' state (resp. action). At step t, the agent observes $\bar{s}_t$ and decides a_t accordingly.

Q8. *With this augmented state $\bar{s}_t$, does the problem reduce to an MDP with no delay? Write the transition probabilities and reward function of this MDP (the rewards might be random).*

Q9. *In the equivalent MDP, consider a deterministic policy π defined by $\pi(\bar{s})$ for all augmented state $\bar{s}$. Write the fixed point equations satisfied by V^π. Propose a RL algorithm to solve them.*