EL2805 Reinforcement Learning

Homework 1

November 8, 2024

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 Dacision Process framework, and to compute (not learn) the optimal control policy.

1.1 Proofreading

A manuscript must be submitted in N hours and has been typed with a known number of mistakes M. Mistakes may be found and corrected through a review. Each review takes one hour to complete and costs an amount $c_1 > 0$. On the k^{th} review, each undetected mistake is found independently with probability p_k. Each undetected mistake left in the manuscript when it is sent to the printer costs an amount $c_2 > 0$. The problem is to decide when to stop reviewing and send the manuscript to the printer.

Q1. *Model the problem as an MDP. Give a precise description of the MDP. Do not try to solve the MDP.*

1.2 Selling your house

You wish to sell your house in Los Angeles. You try to sell it every spring, and start year 1. Due to climate changes, the risk of your house to burn is increasing summer after summer. In year t, the probability that your house disappears due to wildfires is b_t. Each spring you receive offers whose maximum is i.i.d. (across years) and with distribution described by $f(w)$, the probability that the best offer is w, for $w \in \{1, \ldots, W\}$. After selling your house, you place the money and enjoy an interest rate of $r\%$. Your objective is to maximize the average amount of money at the end of year $T > 1$.

Q2. *Model this problem as an MDP (describe the MDP in full detail).*

Q3. *Establish that the optimal policy is threshold-based, i.e., you decide to accept the best offer made in year t if this offer exceeds a threshold.*

Q4. *Provide a general recursive formula satisfied by the thresholds.*

Q5. *Now assume that the best offer distribution is uniform over $[0, 1]$ and that $b_t = b$ for all t. Answer the question c) again in this setting. When T is very large, what are the optimal decisions in the first years?*

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).

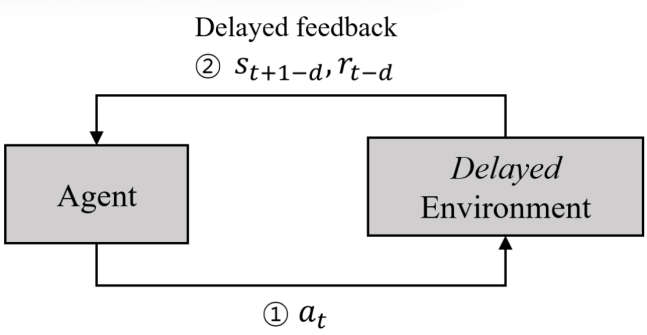

Figure 1: MDP with constant interaction delays. The agent receives delayed information about the state of the system and the rewards.

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.

Q6. *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.

Q7. *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).*

Q8. *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.*