EL2805 Reinforcement Learning

Homework 1

November 6, 2023

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 17, 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 Repair or replace?

You own a bike that can be is several conditions: perfect, worn and broken. You ride your bike every month, and at the beginning of the month you observe its condition and decide what to do. If the bike is broken, you have to either repair it or buy a new bike. If it is worn, you can decide to either keep it as it is, or repair it. When you repair a worn bike, its condition becomes perfect, and when you repair a broken bike, its condition becomes worn. The cost of a new bike is C_b and the cost of repairing the bike is C_r. In one month, the probability that the bike condition degrades is θ (that is, going from perfect to worn, or from worn to broken). You wish to find a strategy that minimizes your expected cost over T months. The condition of the bike at the end of the T-th month does not matter.

a) Model the problem as an MDP, then answer the following question: What is the correct transition matrix? *Note:* The states are indexed as perfect (1), worn (2) and broken (3).

$$P(\text{keep}) = \begin{array}{c} (1) \\ (2) \end{array} \begin{bmatrix} 1-\theta & \theta & 0 \\ 0 & 1-\theta & \theta \end{bmatrix} \quad P(\text{repair}) = \begin{array}{c} (2) \\ (3) \end{array} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

$$P(\text{buy new}) = \begin{array}{c} (2) \\ (3) \end{array} \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}.$$

b) Solve by hand the optimal control problem when there are two decisions ($T = 2$). Then provide an explicit expression of the following quantities as a function of θ, C_r and C_b.

- $u_0^*(\text{Worn}) =$

  with the assumption that repairing costs less than buying a new
  ↓
  $\max(-C_b, -C_r) = -C_r$

  the value of the state worn at the beginning of the first month;

- $a_0^*(\text{Broken}) =$

  Repair (if $-C_r > -C_b$ else buy new)

  the best action if the bike is broken at the beginning of the first month.

What is the best action at the beginning of the first month if the bike is broken for $\theta = 0.5$, $C_b = 9$, $C_r = 6$.

c) Assume that you start with a bike in perfect condition. You decide to never repair nor to buy a new bike. How long does in take in average to get a broken bike? (Here we assume that $T = \infty$).

2 Optimal Stopping

You observe a fair coin being tossed T times. You may stop observing at any time, and when you do you receive as a reward the proportion of heads observed. For example, if the first toss is head, you should stop immediately. Your problem is to identify a stopping rule maximizing the average reward.

a) Model the problem as an MDP. How many states will you use? $1+\frac{T^2+3T}{2}$
Justify your answer and write Bellman's equations.

b) Establish by induction one of the following statement. Which one is true? A
Let $V_t(n)$ denote the maximal average reward if after t tosses, we got n heads.
(A) For all t and n, $V_t(n+1) \geq V_t(n)$
(B) For all t and n, $V_t(n+1) \leq V_t(n)$
(C) For all t and n, $V_t(n+1) = V_t(n)$

c) One of the following policies is optimal. Which one? Justify your choice. Hint: proceed by elimination (justify why 2 of 3 strategies are not optimal). C
(A) After the second toss, stop only if the number of heads reaches $T/2$
(B) Never stop, except when the first toss is head
(C) After t tosses and n observed heads, stop if and only if $n > \frac{t}{2}$

d) The coin is biased, with an unknown bias. We are using an off-policy RL algorithm converging to the optimal policy. The algorithm works with one of the following behavior policies. Which one? B
(A) After t tosses and n observed heads, stop if and only if $n > t/2$
(B) Never stop, i.e., always select the same action