EL2805 Reinforcement Learning

Homework

October 31, 2020

Department 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 15, 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 Machine Replacement

Consider a production machine on a factory floor that can be in three different conditions: perfect, worn and broken. When operating the machine, it has a probability θ of degrading one stage (that is, going from perfect to worn, or from worn to broken). The factory owner can choose to replace the machine at a cost R. If it is broken, then he acquires a cost c for not being able to produce new products, and if it is worn, he acquires a cost $c/2$ for producing imperfect items (at each time-step). He wants to find an optimal policy for T time-steps that minimizes his expenses. Assume that the cost at the end of the horizon is zero, regardless of state.

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).

$$P(\text{replace}) = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \text{ and } P(\text{continue}) = \begin{bmatrix} 1-\theta & \theta & 0 \\ 0 & 1-\theta & \theta \\ 0 & 0 & 1 \end{bmatrix}.$$

b) For $\theta = 0.5$, $R = 8$, $c = 6$, $T = 2$, solve the MDP *by hand*. That is, compute the optimal cost-to-go and the optimal policy. Then answer the following questions:

- $u_0^*(\text{Worn}) =$ -15/2
- $a_0^*(\text{Broken}) =$ replace

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? $\underline{T^2+1}$
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)**[1] One of the following policies is optimal. Which one? Justify your choice._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? _A____
(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

[1] A difficult question – not qualifying to pass the HW.