Reinforcement Learning (EL2805) HW 1

Lab Report - EL2320, Autumn 2023

Oscar Eriksson | Philipp Ahrendt
oscer@kth.se | pcah@kth.se

November 2023

Part 1

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)

We have 3 actions $\mathcal{A} = \{K, R, N\}$ (keep, repair, buy new) and 3 states, $\mathcal{S} = \{P, W, B\}$ (perfect, worn broken). The cost of a new bike is C_b and the cost of repairing a bike is C_r. Every time-step there is a θ chance that the bike's condition deprecates if it is not already broken. We thus define the 3 transition probability matrices

$$P^K = \begin{bmatrix} 1-\theta & \theta & 0 \\ 0 & 1-\theta & \theta \\ 0 & 0 & 1 \end{bmatrix}, \quad P^R = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad P^N = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

with the states P, W, B corresponding to the first, second and third rows/columns respectively. The cost incurred by choosing any action from any given state is described by the following 3 reward matrices

$$R^K = \begin{bmatrix} 0 \\ 0 \\ -\infty \end{bmatrix}, \quad R^R = \begin{bmatrix} -C_r \\ -C_r \\ -C_r \end{bmatrix}, \quad R^N = \begin{bmatrix} -C_b \\ -C_b \\ -C_b \end{bmatrix}$$

(one may not keep a broken bike).

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.

We solve the recursion $u_t = \max_a[R^a + P^a u_{t+1}]$. We start with $u_T = u_2 = (0, 0, 0)$ reflecting the fact that we don't care what the condition is by the end of the second month (u_2 corresponds to the beginning of the third month). We then take

$$u_1 = \max_a[R^a] = \begin{bmatrix} 0 \\ 0 \\ \max(-C_b, -C_r) \end{bmatrix}, \quad a_1^\star = \begin{bmatrix} K \\ K \\ R \text{ if } -C_r > -C_b \text{ else } N \end{bmatrix}$$

We find that we only want to spend cash if we have no other choice, but then we choose the cheap option. Hence we don't even consider repairing a worn bike. Finally,

$$u_0 = \max_a[R^a + P^a u_1] = \begin{bmatrix} 0 \\ \theta\max(-C_b, -C_r) \\ \max(-C_b, -C_r) \end{bmatrix}, \quad a_0^\star = \begin{bmatrix} K \\ K \\ R \text{ if } -C_r > -C_b \text{ else } N \end{bmatrix}$$

We find that the same logic applies as above, but we get the sense that an even longer horizon would make the policy less straight forward. $(u_0^\star)_W = \theta\max(-C_b, -C_r) = -3$, $(a_0^\star)_B = (R \text{ if } -C_r > -C_b \text{ else } N) = R$

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 $T = \infty$).

We simply take the expected value of the time until the bike is worn (starting in perfect condition), and add the expected value of the time until the bike is broken (starting in worn condition). Since the probability of a bike diminishing in quality is θ regardless of starting position, these expected value will be equal. We write the probability of the bike deprecating after $k+1$ time-steps

$$p_{\text{dec}}(k+1;\theta) = (1-\theta)^k \theta$$

and hence the expected value of k_{dec} is

$$\begin{aligned}
\mathbb{E}[k_{\text{dec}}] &= \sum_{k=0}^{\infty}(k+1)(1-\theta)^k\theta = \sum_{k=0}^{\infty}\frac{\partial}{\partial\theta}\left[-(1-\theta)^{k+1}\right]\theta \\
&= -\frac{\partial}{\partial\theta}\left[\sum_{k=0}^{\infty}(1-\theta)^{k+1}\right]\theta = -\frac{\partial}{\partial\theta}\left[\sum_{k=1}^{\infty}(1-\theta)^{k}\right]\theta \\
&= -\frac{\partial}{\partial\theta}\left[\frac{1}{1-(1-\theta)}\right]\theta = -\frac{\partial}{\partial\theta}\left[\frac{1}{\theta}\right]\theta = \frac{1}{\theta} \\
\mathbb{E}[k_{\text{break}}] &= 2\mathbb{E}[k_{\text{dec}}] = \frac{2}{\theta}
\end{aligned}$$

In our case $\theta = 1/2$, and as such we find that $\mathbb{E}[k_{\text{dec}}] = 2$. This means that $\mathbb{E}[k_{\text{break}}] = 2\mathbb{E}[k_{\text{dec}}] = 4$ □.

Part 2

a) Model the problem as an MDP. How many states will you use? Justify your answer and write Bellman's equations.

The states space can be written as $\mathcal{S} = \{n_1, n_2, ...n_T\} \times \{t_1, t_2, ..., t_T\} \cup \{e\}$ (n_k corresponding to having k heads, t_k corresponding to being at the k:th timestep and e being a self-looping exit state to which we move when we wish to stop observing). We then define the actions $\mathcal{A} = \{C, E\}$ (continue, end). Although normally this would result in a state space of $T^2 + 1$ states but since it is not possible to have a greater number of heads than timesteps. This results in

$$\begin{aligned}
\text{Number of used states} &= 1 + 2T + \sum_{i=1}^{T-1} T - i = 1 + 2T + \sum_{i=1}^{T} T - i = 1 + 2T + T^2 + \sum_{i=1}^{T} -i = 1 + 2T + T^2 - \frac{T(T+1)}{2} \\
&= 1 + 2T + \frac{2T^2 - T^2 - T}{2} = 1 + 2T + \frac{T^2 - T}{2} = 1 + \frac{4T + T(T-1)}{2} \\
&= 1 + \frac{T^2 + 3T}{2}
\end{aligned}$$

We state the transition probabilities below:

$$\begin{aligned}
&P^C(s_{k+1} = (n_{i+1}, t_{k+1}) \mid s_k = (n_i, t_k)) = 1/2, a = C, \ \forall k = 1, ..., T, \ i = 0, ..., T \\
&P^C(s_{k+1} = (n_i, t_{k+1}) \mid s_k = (n_i, t_k)) = 1/2, a = C, \ \forall k = 1, ..., T, \ i = 0, ..., T \\
&P^E(s_{k+1} = e) = 1, \ \forall k = 1, ..., T
\end{aligned}$$

and let any unstated transition have probability 0. We then define the rewards below

$$\begin{aligned}
R^C(s_k = (n_i, t_k)) &= 0, \ \forall k = 1, ..., T, \ i = 0, ..., T \\
R^E(s_k = (n_i, t_k)) &= i/k, \ \forall k = 1, ..., T, \ i = 0, ..., T
\end{aligned}$$

and let any unstated reward have probability 0. Bellman's equations are simply

$$\begin{aligned}
u_{T+1} &= 0 \\
u_k &= \max_a [R^a + P^a u_{k+1}], \quad \forall k = 1, ..., T \\
u_k &= \max_a [R(n_i, t_k), \frac{u_{k+1}(n_{i+1}, t_{k+1}) + u_{k+1}(n_i, t_{k+1})}{2}] \\
a_k &= \text{argmax}_a [R^a + P^a u_{k+1}], \quad \forall k = 1, ..., T
\end{aligned}$$

b) Establish by induction one of the following statement. Which one is true? Let $V_t(n)$ denote the maximal average reward if after t tosses, we got n heads

Let's write the Bellman equation of 2.a in this notation

$$V_t(n) = \max_a [R(n), \frac{V_{t+1}(n+1) + V_{t+1}(n)}{2}]$$

From a first intuition we assume answer A being correct. With this assumption we can show For $t = T$, it follows that

$$V_T(n) = \max_a [R(n)] = \frac{n}{T}$$

And thus

$$V_T(n+1) = \max_a[R(n+1)] = \frac{n+1}{T} \geq \frac{n}{T} = \max_a[R(n)] = V_T(n)$$

We can see that the assumption holds for $t = T$.
For $t = 0, ..., T-1$:

$$V_t(n+1) = \max_a[R(n+1), \frac{V_{t+1}(n+2) + V_{t+1}(n+1)}{2}] \geq \max_a[R(n), \frac{V_{t+1}(n+1) + V_{t+1}(n)}{2}] = V_t(n)$$

In the case of $R(n) > \frac{V_{t+1}(n+1)+V_{t+1}(n)}{2}$

$$V_t(n+1) = R(n+1) = \frac{n+1}{t} \geq \frac{n}{t} = R(n) = V_t(n)$$

In the other case:

$$\begin{aligned} V_t(n+1) = \frac{V_{t+1}(n+2) + V_{t+1}(n+1)}{2} &\geq \frac{V_{t+1}(n+1) + V_{t+1}(n)}{2} = V_t(n) \\ V_{t+1}(n+2) &\geq V_{t+1}(n) \\ \Rightarrow V_t(n+1) &\geq V_t(n) \end{aligned}$$

Since this holds for any $t = 1, ..., T$ we have shown that assumption A is indeed correct

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).

When we let our markov process elapse without interruption (no E action) we see that it behaves like a binomial distribution; the likelihood of n heads after t timesteps is described by

$$p_t(n) = \binom{t}{n} (1-\theta)^{t-n}\theta^n$$

for $\theta = 0.5$. If we want the expected value of n/t we need to sum over the terms $i_t(n)$ for all n

$$\begin{aligned} i_t(n) &= \frac{n}{t}\binom{t}{n}(1-\theta)^{t-n}\theta^n = \frac{n}{t}\frac{t!}{n!(t-n)!}(1-\theta)^{t-n}\theta^n \\ &= \frac{(t-1)!}{(n-1)!(t-n)!}(1-\theta)^{t-n}\theta^n = \theta\binom{t-1}{n-1}(1-\theta)^{t-n}\theta^{n-1} = \theta p_{t-1}(n-1) \end{aligned}$$

We recognize that summing all $\sum p_{t-1}(n-1) = 1$ and as such $\mathbb{E}[n/t] = \sum i_t(n) = \theta = 1/2$. The conclusion we draw is that if our stopping policy is to never stop, the expected reward is 1/2 (given that we actually cash out). We therefore expect an optimal stopping policy to yield an expected reward/value greater than 1/2.

(A): This is not optimal. With this policy the stopping rule will yield a reward of $n/T = T/2T = 1/2$, which does not perform better than having no policy at all. We can easily improve this policy by having the criteria that we stop observing when $n/t > \mathbb{E}[n/t] = 1/2$.

(B): This is not optimal. Not stopping at any point yields an expected reward of 1/2. We can easily improve this policy by having the criteria that we stop observing when $n/t > \mathbb{E}[n/t] = 1/2$.

(C): If not (A) nor (B) is optimal, then this should be. We already stated that stopping when $n/t > \mathbb{E}[n/t] = 1/2$ is a better policy than (A) and (B), but we recognize that we do not need to prove that it is optimal ourselves.

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?

Since we rely on exploring the space in off-policy RL and with answer A we would not fully explore the space if the coin is for example biased towards head and stop too early.
For instance if the coin is heavily biased for heads (like 0.99), then with policy (A) it would take hundreds of trials to figure out the bias of the coin, since it would terminate after one step with probability 0.99. That way, even after hundreds of trials, if our first toss shows tails and our second toss shows heads then the policy would likely terminate early (suboptimally) instead of awaiting the expected 0.99 it would receive from not not terminating at all.
The policy B would fully explore the space in an episode giving us more information about the bias of the coin so that we can update our policy in the next step. We therefore consider policy B to be the desired policy.

References