EL2805 - Homework 1

Lea Keller, Jannik Wagner

November 15, 2022

1 Repair or replace?

a)

$$S = \{perfect, worn, broken\} = \{p, w, b\}$$

$$A_p = \{keep\} = \{k\}$$
$$A_w = \{keep, repair, buy\} = \{k, r, b\}$$
$$A_b = \{repair, buy\} = \{r, b\}$$

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

$$r(p, k) = r(w, k) = 0$$
$$r(w, r) = r(b, r) = -C_r$$
$$r(w, b) = r(b, b) = -C_b$$

b)

$$u_1(p) = \max\{r(p, k)\} = 0$$
$$u_1(w) = \max\{r(w, k), r(w, r), r(w, b)\} = \max\{0, -6, -8\} = 0$$
$$u_1(b) = \max\{r(b, r), r(b, b)\} = \max\{-6, -8\} = -6$$

$$u_0(p) = \max\{r(p, k) + (1-\theta)u_1(p) + \theta u_1(w)\} = 0$$
$$a_0(p) = k$$

$$u_0(w) = \max\{r(w,k)+(1-\theta)u_1(w)+\theta u_1(b), r(w,r)+(1-\theta)u_1(p)+\theta u_1(w),$$

$$r(w,b)+(1-\theta)u_1(p)+\theta u_1(w) = \max\{-3,-6,-8\} = -3$$

$$a_0(w) = k$$

$$u_0(b) = \max\{r(b,r)+(1-\theta)u_1(w)+\theta u_1(b), r(b,b)+(1-\theta)u_1(p)+\theta u_1(w)$$

$$= \max\{-9,-8\} = -8$$

$$a_0(b) = b$$

c)

We where probably supposed to solve this with some of the tools that were given to us in the lecture. However, it felt more intuitive to solve it the following way.

Let X_t be 1 if the condition degrades at month t and 0 otherwise. Then $P(X_t = 1) = \theta$ and $P(X_t = 0) = 1 - \theta$.

Let $Y_t = \sum_{i=1}^{t} X_i$ the number of times the bikes condition degraded up to month t.

Let $T_k = \min\{t : Y_t = k\}$ be the first month at which the bike's condition has degraded k times.

Then

$$P(T_k = n) = \theta^k (1-\theta)^{n-k} \binom{n-1}{k-1}$$

since we can choose the position of k-1 0s among the n-1 first throws, however, the nth throw must be a 1.

We are interested in $\mathbb{E}[T_2]$.

Let $\alpha = 1 - \theta$.

$$\begin{aligned}
\mathbb{E}[T_2] &= \sum_{t=0}^{\infty} P(T_2 = t)t \\
&= \sum_{t=0}^{\infty} \theta^2 \alpha^{t-2}(t-1)t \\
&= \theta^2 \sum_{t=0}^{\infty} \alpha^{t-2}(t-1)t \\
&= \theta^2 \frac{2}{(1-\alpha)^3} \\
&= \theta^2 \frac{2}{\theta^3} \\
&= \frac{2}{\theta}
\end{aligned}$$

2 Optimal Stopping

a)

States: number of heads and stopped state.

$$S = \{0, ..., T, E\}$$

Actions:

$$A = \{continue, end\} = \{c, e\}$$

Terminal reward

$$r_T(n) = \frac{n}{T}$$

$$r_T(E) = 0$$

$t < T$

$$r_t(n, c) = 0$$

$$r_t(n, e) = \frac{n}{t}$$

Probabilities:

$$p(n|n, c) = p(n+1|n, c) = \frac{1}{2}$$

$$p(E|n, e) = p(E|E, e) = 1$$

All other 0.

Bellman equation:

$$V_t(n) = \max\{\frac{n}{t}, \frac{1}{2}V_{t+1}(n+1) + \frac{1}{2}V_{t+1}(n)\}$$

Equivalently, we could include the time t in the state, so that we have states (t, n) for all $t \geq n$ instead of n. In that case

$$p((t+1, n')|a, (t, n)) = p(n'|a, n)$$

and

$$r((n, t), a) = r_t(n, a)$$

However, for simplicity we decided to model t with the time parameter of r_t, since we are considering a finite horizon MDP.

b)

(A) is correct

Theorem 1. *For all t and n, $V_t(n+1) \geq V_t(n)$*

Proof. Proof by induction.

Start of induction: $t = T$

$$V_T(n+1) = \frac{n+1}{T} \geq \frac{n}{T} = V_T(n)$$

Induction step: t to $t-1$

We need to show that

$$V_{t-1}(n+1) \geq V_{t-1}(n)$$

Bellman's equation gives us:

$$V_{t-1}(n+1) = \max\{\frac{n+1}{t-1}, \frac{1}{2}V_t(n+2) + \frac{1}{2}V_t(n+1)\}$$

$$V_{t-1}(n) = \max\{\frac{n}{t-1}, \frac{1}{2}V_t(n+1) + \frac{1}{2}V_t(n)\}$$

Either $\frac{n}{t-1} \geq \frac{1}{2}V_t(n+1) + \frac{1}{2}V_t(n)$

in which case

$$V_{t-1}(n) = \frac{n}{t-1} \leq \frac{n+1}{t-1} \leq V_{t-1}(n+1)$$

Or $\frac{n}{t-1} \leq \frac{1}{2}V_t(n+1) + \frac{1}{2}V_t(n)$

in which case

$$V_{t-1}(n) = \frac{1}{2}V_t(n+1) + \frac{1}{2}V_t(n) \overset{IH}{\leq} \frac{1}{2}V_t(n+2) + \frac{1}{2}V_t(n+1) \leq V_{t-1}(n+1)$$

□

The same argument can also be made for $V_t(n+1) > V_t(n)$.

c)

The policy (C) *After t tosses and n observed heads, stop if and only if* $n > \frac{t}{2}$ is optimal.

The policy (A) *After the second toss, stop only if the number of heads reaches* $\frac{T}{2}$ is not optimal.

Consider the case that you get heads in the first throw, i.e., $t = n = 1$. Under a policy that stops in this situation π', the value would be $V_1^{\pi'}(1) = 1$. In consequence,

$$\begin{aligned} V_1^{\pi_A}(1) &= \frac{1}{2}V_2^{\pi_A}(2) + \frac{1}{2}V_2^{\pi_A}(1) \\ &< \frac{1}{2}V_2^{\pi_A}(2) + \frac{1}{2}V_2^{\pi_A}(2) = V_2^{\pi_A}(2) \leq 1 = V_1^{\pi'}(1) \leq V_1^{*}(1) \end{aligned}$$

I.e., the policy is not optimal.

The policy (B) *Never stop, except when the first toss is head* is not optimal.

Consider the following example: $T = 8$, Assume we are at $t = 4$ and have seen 0, 1, 1, 1 so far. $V_4^{\pi'}(3) = \frac{3}{4}$ would be the value under a policy that stops here (as the optimal policy would). However, this policy will continue until $t = T$ and thus on average get 2 more heads. In consequence,

$$V_4^{\pi_B}(3) = V_8^{\pi_B}(5) = \frac{5}{8} < \frac{3}{4} = V_4^{\pi'}(3) \leq V_4^{*}(3)$$

I.e., the policy is not optimal.

d)

(B) is correct.

The policy used in the algorithm is *Never stop, i.e., always select the same action.*

With this policy, the algorithm can estimate the value of all states and converge to an optimal policy which can handle all states.

Policy (A) *After t tosses and n observed heads, stop if and only if $n > t/2$* is an never explores taking action continue in states with $n > \frac{t}{2}$, the learning algorithm can thus not reliably estimate their value and thus does not converge to an optimal policy.