Homework 1: Solutions

ORF 418 - Fall 2025

version: August 25, 2025

PLEASE REVISE

1. PLEASE CHECK.

a. Recall from the lecture notes that

$$A = \begin{bmatrix} 1 & h & 0 & 0 \\ 0 & 1 & ah & 0 \\ 0 & 0 & 1 & h \\ 0 & 0 & bh & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ hL \\ 0 \\ h \end{bmatrix}.$$

Using the values from the statement, we therefore obtain

$$A = \begin{bmatrix} 1 & 0.01 & 0 & 0 \\ 0 & 1 & 0.001 & 0 \\ 0 & 0 & 1 & 0.01 \\ 0 & 0 & 0.051 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0.02 \\ 0 \\ 0.01 \end{bmatrix}.$$

b. The general expression of the Riccati equation reads

$$V = M + \rho A^T V A - \rho^2 A^T V B \left(N + \rho B^T V B \right)^{-1} B^T V A.$$

The quantity $N + \rho B^T V B$ is here a scalar, so the Riccati equation can be rewritten in a more compact form. Since B has only two non-zero entries, observe that

$$B^T V B = (hL)^2 V_{22} + h^2 L V_{24} + h^2 L V_{42} + h^2 V_{44} = h^2 \left[L^2 V_{22} + 2 L V_{24} + V_{44} \right],$$

as V is symmetric by assumption. Thus,

$$V = M + \rho A^T V A - \left(\frac{\rho}{h} \right)^2 \frac{A^T V B B^T V A}{h^{-2} + \rho \left(L^2 V_{22} + 2 L V_{24} + V_{44} \right)}.$$

Moreover, one has $BB^T = h^2 G$, where

$$G = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & L^2 & 0 & L \\ 0 & 0 & 0 & 0 \\ 0 & L & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 2 \\ 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 1 \end{bmatrix}.$$

This finally gives the semi-explicit expression,

$$V = M + \rho A^T V A - \rho^2 \frac{A^T V G V A}{h^{-2} + \rho\,(L^2\,V_{22} + 2LV_{24} + V_{44})}$$

$$= M + 0.81\,A^T V A - 0.81^2 \frac{A^T V G V A}{10^4 + 0.81(4\,V_{22} + 4V_{24} + V_{44})}.$$

c. Please refer to the Jupyter notebook. The solution and LQR gains matrix are respectively

$$V = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 5.37 & 0.23 \\ 0 & 0 & 0.23 & 0.02 \end{bmatrix}, \quad F = [\ 0\ ,\ 0\ ,\ 0.002\ ,\ 0\ ].$$

d. The eigenvalues of $A - BF$ are $(0.90, 0.90, 0.88, 0.92)$ (cf. Jupyter notebook). As all of them are less than 1, the dynamics is stable. Moreover, the linearized inverted pendulum is always controllable as long as $a \neq 0$, $b \neq 0$ and $a \neq Lb$, which is the case here. See Example 2.4 from the lecture notes for further details.

2. **a.** The following trick proves useful: we add a fourth arm, say $a = 4$, and consider the random decision

$$b_t = \begin{cases} 1, 2 \text{ or } 3, & \text{each with probability } \frac{\epsilon}{3} = 0.05, \\ 4, & \text{with probability } 1 - \epsilon = 0.85. \end{cases}$$

If arm 4 is chosen, we are always greedy and pick the most profitable arm based on our estimates of $(q(a))_{a=1}^3$, i.e. $a_t = a_t^* = \underset{a=1,2,3}{\operatorname{argmax}}\ \hat{q}_T(a)$. Otherwise, if $b_t < 4$, we simply set $a_t = b_t$. Summarizing, we have

$$a_t = \begin{cases} b_t, & \text{if } b_t < 4, \\ a_t^*, & \text{if } b_t = 4. \end{cases}$$

Next, fix $a \in \{1, 2, 3\}$ and observe that

$$\{a_t = a\} = \{b_t = a\} \sqcup \{a_t^* = a, b_t = 4\}, \quad (\star)$$

where $\sqcup$ indicates that the two events on the right-hand side are disjoint. Hence, for any $a \in \{1, 2, 3\}$,

$$\mathbb{P}(a_t = a) \geq \mathbb{P}(b_t = a) = \frac{\epsilon}{3} = 0.05 > 0.$$

In view of $(\star)$, $\mathbb{1}_{\{a_t = a\}} \geq \mathbb{1}_{\{b_t = a\}}$, and hence

$$\frac{N_T(a)}{T} = \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}_{\{a_t = a\}} \geq \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}_{\{b_t = a\}}.$$

Since b_t's are i.i.d., we use the law of large numbers (LLN) to conclude that Taking the infimum limit on both sides (note that $\frac{N_T(a)}{T}$ need not be $\mathbb{P}-$a.s. convergent) leads to

$$\liminf_{T\to\infty} \frac{N_T(a)}{T} \geq \liminf_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_{\{b_t=a\}} = \mathbb{E}[\mathbb{1}_{\{b_1=a\}}] = \frac{\epsilon}{3} = 0.05 > 0.$$

Note that the sequence $(a_t)_{t\geq 1}$ is not i.i.d., and we could not immediately use the LLN. In fact, this is the reason to introduce the sequence b_t.

□

b. By part **a.**, $\lim_{T\to\infty} N_T(a) = \infty$. We again aim to use the LLN, but we cannot apply it directly. Indeed, the variables $(r_t \mathbb{1}_{\{a_t=a\}})_{t\geq 1}$ are *not* independent nor identically distributed. To circumvent this issue, fix $a \in \{1,2,3\}$ and consider the *arrival times*,

$$\begin{aligned}\tau_1 &= \inf\{t \geq 1 \,|\, a_t = a\},\\ \tau_k &= \inf\{t > \tau_{k-1} \,|\, a_t = a\}, \quad k \geq 2.\end{aligned}$$

We omit the dependence of τ_k on a for simplicity. As $\lim_{T\to\infty} N_T(a) = \infty$ implies $\tau_k < \infty$ with probability one, this allows us to properly define $z_k := r_{\tau_k}$, the $k-th$ reward received from arm a. Then, for $\tau_k \leq T < \tau_{k+1}$, we have $N_T(a) = k$ and for all such T,

$$\widehat{q}_T(a) = \frac{1}{N_T(a)} \sum_{t=1}^{T} r_t \, \mathbb{1}_{\{a_t=a\}} = \frac{1}{k}\sum_{t=1}^{k} z_t.$$

Therefore, by LLN,

$$\lim_{T\to\infty} \widehat{q}_T(a) = \lim_{k\to\infty} \frac{1}{k}\sum_{t=1}^{k} z_t \;=\; \mathbb{E}[z_1],$$

since the rewards (z_k) are independent and generated from the same (conditional) distribution, namely the one of $r_t|\, a_t = a$. Therefore, $\mathbb{E}[z_1] = q(a)$ as claimed.

□

c. For any $a \in \{1,2,3\}$, the decomposition $(\star)$ from part **a.** implies that

$$\frac{N_T(a)}{T} = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_{\{b_t=a\}} + \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_{\{a_t^*=a, b_t=4\}}.$$

We now use the fact that the agent will eventually find out that arm 3 is the most profitable. Having proven in part **b.** that $\lim_{T\to\infty} \widehat{q}_T(a) = q(a)$, there thus exists an $\mathbb{P}-$a.s. finite random time θ such that $a_t^* = 3$ for all $t \geq \theta$. As we picked $a \in \{1,2\}$, the variable $\mathbb{1}_{\{a_t^*=a, b_t=4\}}$ will vanish for all $t \geq \theta$. Hence, with probability one,

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_{\{a_t^*=a, b_t=4\}} = \frac{1}{T}\sum_{t=1}^{\theta\wedge T} \underbrace{\mathbb{1}_{\{a_t^*=a, b_t=4\}}}_{\leq 1} \leq \frac{\theta\wedge T}{T} \;\overset{T\uparrow\infty}{\longrightarrow}\; 0.$$

We conclude that

$$\lim_{T\to\infty} \frac{N_T(a)}{T} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}_{\{b_t=a\}} = \mathbb{P}(b_1 = a) = \frac{\epsilon}{3} = 0.05.$$

Finally, for $a = 3$, we simply remark that $\frac{N_T(3)}{T} = 1 - \frac{N_T(1)}{T} - \frac{N_T(2)}{T}$, leading to $\lim_{T\to\infty} \frac{N_T(3)}{T} = 1 - \epsilon = 0.9$ by linearity of the limit.

□

d. Combining the above observations, we obtain

$$\begin{aligned}
\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} r_t &= \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} r_t \underbrace{\sum_{a=1}^{3} \mathbb{1}_{\{a_t=a\}}}_{=1} \\
&= \sum_{a=1}^{3} \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} r_t \mathbb{1}_{\{a_t=a\}} \\
&= \sum_{a=1}^{3} \left(\lim_{T\to\infty} \frac{N_T(a)}{T} \right) \left(\lim_{T\to\infty} \frac{1}{N_T(a)} \sum_{t=1}^{T} r_t \mathbb{1}_{\{a_t=a\}} \right) \\
&= \sum_{a=1}^{3} p(a)\, q(a) \quad (\textbf{b.} + \textbf{c.}) \\
&= \frac{\epsilon}{3} [q(1) + q(2)] + (1 - \frac{2}{3}\epsilon) q(3) \\
&= 9.15.
\end{aligned}$$

□

3. This is done in the above.