ORF 418 - Optimal Learning - Fall 2024
Instructor: Emma Hubert

Old exams are provide for study purposes and they are *not* representative of this year's exam because:
1. This year's exam also covers parts of Chapter 4 (Kalman Filter).
2. These exams were prepared by a different instructor.

October 23, in class (8:30am – 9:50am)
Please carefully justify all your answers.

Exercise 1. A three-armed bandit problem. [30pt]

We consider a multi-armed bandit problem with **three arms**, $a \in \{1, 2, 3\}$. The random rewards from arm $a \in \{1, 2, 3\}$ are distributed according to a **Bernoulli** distribution with probability of success $a/4$. All the rewards are assumed to be independent.

We first **explore each arm** N **times** and obtain the rewards $\{r_k^{(a)}\}_{k=1,\dots N}$. We use the standard estimate

$$\widehat{q}(a) := \frac{1}{N} \sum_{k=1}^{N} r_k^{(a)}, \qquad a \in \{1, 2, 3\}.$$

to compute $a^\star := \arg\max_{a\in\{1,2,3\}} \widehat{q}(a)$. We further assume that, if there are more than one maximisers, we choose the **smallest**. We then **exploit** by using **only** $a^\star$ and receive independent rewards $(r_1, r_2, \dots)$. The limiting reward is

$$J(a^\star) := \lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{T} r_k.$$

a. **[5pt]** Assume $N = 1$ and $(r_1^{(1)}, r_1^{(2)}, r_1^{(3)}) = (0, 1, 1)$. Compute $a^\star$ and $J(a^\star)$.

b. **[5pt]** Compute $J(a^\star)$ for $a^\star = 1$ and $a^\star = 3$.

c. **[10pt]** Still assuming that $N = 1$, compute the distribution of $a^\star$.

d. **[10pt]** We now assume $N \longrightarrow +\infty$. Deduce $a^\star$ and compute the limiting reward if, instead of using only $a^\star$, we implement an ε-greedy algorithm with $\varepsilon = 0.2$ (briefly describe the algorithm).

Exercise 2. A (linear quadratic) control problem. [25pt]

Consider the following one-dimensional control problem with state dynamics

$$x_{k+2} - 0.5x_k = bu_k, \ k \in \mathbb{N},$$

for $(x_0, x_1) \in \mathbb{R}^2$ and $b \in \mathbb{R}$. The cost function (to be minimised) is given by

$$J(x_0, x_1, u) := \sum_{k=0}^{\infty} \left(x_k^2 + u_k^2\right).$$

a. **[5pt]** Show by mathematical induction that, if $b = 0$, then $x_{2k} = 0.5^k x_0$ and $x_{2k+1} = 0.5^k x_1$, for all $k \in \mathbb{N}$.

b. **[5pt]** Formulate the previous control problem as a standard linear-quadratic problem in dimension $d = 2$ (write A, B, M, N and ρ). Give a sufficient and necessary condition for the system to be controllable.

c. **[5pt]** Define the value function v corresponding to this control problem. Is the value function finite?

d. **[10pt]** State the general Dynamic Programming Equation and use it to show that the value function satisfies:

$$v(x_0, x_1) = x_0^2 + \inf_{u\in\mathbb{R}} \left\{u^2 + v(\widetilde{x})\right\}, \qquad \forall (x_0, x_1) \in \mathbb{R}^2, \tag{2.1}$$

for $\widetilde{x} \in \mathbb{R}^2$ to be determined ($\widetilde{x}$ is allowed to depend on x_0, x_1, u and the parameters of the model).

Exercise 3. Bayesian estimation. [25pt]

Let $\theta \in (0,1)$ be the **unknown** proportion of individuals infected by a virus in a given population. To estimate θ, we test 100 individuals and denote by Y the number of infected individuals among them. We observe $Y = 10$.

a. **[10pt]** Write the conditional probability $\mathbb{P}(Y = y|\theta)$ for any $y \in \{0, \dots, 100\}$ and compute the maximum likelihood estimator corresponding to the observation $Y = 10$.

b. **[10pt]** We now assume that the prior distribution for θ is a Beta distribution with parameters $a > 0$ and $b > 0$. Show that the posterior distribution for θ given the observation $Y = 10$ is again a Beta distribution, but with parameters $(a + 10, b + 90)$.

c. **[5pt]** Compute the Bayesian estimator (for a quadratic loss) and compare it to the MLE estimator.

Hint. The Beta distribution with parameters $a, b > 0$ is characterised by the following density,

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1}, \qquad x \in (0,1),$$

where Γ is the gamma function. Moreover, its mean is equal to $a/(a+b)$.

Exercise 4. Another control problem. [20pt]

Consider the following one-dimensional ($d = \ell = 1$) control problem with finite time horizon $n \in \mathbb{N}^\star$ and state dynamics

$$x_{k+1} = x_k + bu_k, \quad k \in \{0, \dots, n-1\},$$

for some $b \in \mathbb{R}$ and $x_0 \in \mathbb{R}$ fixed. We denote by $\bar{u}_k := (u_k, \dots, u_{n-1})$ the sequence of controls starting from step $k \in \{0, \dots, n-1\}$. The **reward** function when starting at step $k \in \{0, \dots, n\}$ from state $x \in \mathbb{R}$ is defined by

$$J\big(k, x, \bar{u}_k\big) := \rho^{n-k} x_n - \sum_{i=k}^{n-1} \rho^{i-k} u_i^2, \text{ for } k \in \{0, \dots, n-1\}, \text{ and } J(n, x) = x,$$

for some **discount factor** $\rho \in (0,1)$. Finally, we define the value function of this **maximisation** problem, when starting from state $x \in \mathbb{R}$ at step $k \in \{0, \dots, n-1\}$, by

$$v(k, x) := \sup_{\bar{u}_k} J(k, x, \bar{u}_k).$$

a. **[5pt]** Give a sufficient and necessary condition for the system to be controllable.

b. **[5pt]** Show that for all $k \in \{0, \dots, n-1\}$,

$$J\big(k, x, \bar{u}_k\big) = -u_k^2 + \rho J\big(k+1, x + bu_k, \bar{u}_{k+1}\big).$$

c. **[10pt]** Deduce the dynamic programming equation satisfied by the value function $v(k, x)$ for $k \in \{0, \dots, n-1\}$ and $x \in \mathbb{R}$. Specify $v(n, x)$.