ORF 418 - Optimal Learning - Fall 2023
Emma Hubert

Midterm

October 25, in class (8:30am – 9:50am)

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

We consider a multi-armed bandit problem with two arms, $a \in \{1, 2\}$. The reward from arm $a = 1$ is uniform on the interval $[0, 1]$, while the reward from arm $a = 2$ is a Bernoulli with probability $p = 1/3$ of success. The rewards are independent.

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

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

and compute $a^\star := \arg\max_{a\in\{1,2\}} \widehat{q}(a)$. We then **exploit** by using **only** $a^\star$ and receive independent rewards $(r_1, r_2, \dots)$. The limiting reward is defined as

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

a. **[10pt]** Compute the limiting reward $J(a^\star)$ for both $a^\star = 1$ and $a^\star = 2$.

b. **[10pt]** Let $N = 1$, *i.e.* we try **each arm only once**. Compute the distribution of $a^\star$ as well as $\mathbb{E}[J(a^\star)]$.

c. **[10pt]** Consider now the case $N = 2$, *i.e.* we try **each arm twice**. Compute the distribution of $a^\star$ and deduce $\mathbb{E}[J(a^\star)]$.

Exercise 2. [35pt] A Linear-Quadratic Problem.

Consider the following one-dimensional ($d = \ell = 1$) **Linear-Quadratic control problem** with state dynamics

$$x_{k+1} = ax_k + bu_k,\ k \geq 0,$$

for some $a \in \mathbb{R}$, $b \in \mathbb{R}$ and $x_0 \in \mathbb{R}$ arbitrary fixed, and cost function (to be minimised)

$$J(x_0, u) := \sum_{k=0}^{\infty} \frac{1}{2^k} \Big(2x_k^2 + 3u_k^2 \Big).$$

a. **[5pt]** For which conditions on the parameter $b \in \mathbb{R}$ is the system controllable?

b. **[10pt]** Define the value function v of this LQ control problem. For which conditions on the parameters $a \in \mathbb{R}$, $b \in \mathbb{R}$ and $x_0 \in \mathbb{R}$ is the value function finite?

c. **[10pt]** In the following, we assume $b \neq 0$. State the general Dynamic Programming Equation and use it to show that the value function satisfies:

$$v(x) = 2x^2 + \inf_{u\in\mathbb{R}} \left\{ 3u^2 + \frac{1}{2} v(ax + bu) \right\}, \qquad \forall x \in \mathbb{R}. \tag{0.1}$$

d. **[10pt]** Assume that the value function is of the form $v(x) = Vx^2$ for all $x \in \mathbb{R}$ and $V \in \mathbb{R}$ to be determined. Use the dynamic programming equation (or the corresponding Riccati equation) to show that the optimal control is given by

$$u_k^\star = \frac{-abV}{6 + Vb^2}\, x_k, \quad k \geq 0,$$

where $V \in \mathbb{R}$ is an appropriate solution to the following quadratic equation (you do not need to solve it):

$$b^2V^2 + (6 - 2b^2 - 3a^2)V - 12 = 0.$$

Exercise 3. [35pt] Bayesian estimation.

Let θ be the proportion of students enrolled in ORF 418 who would like to visit Paris (France). We assume that θ is unknown and we want to estimate it by asking students one by one. We will denote by $Y_k = 1$ if student k wants to visit Paris, and $Y_k = 0$ otherwise. We assume that the prior distribution for θ is uniform on $[0, 1]$.

a. **[5pt]** We ask a first student and observe $Y_1 = 1$. Compute the posterior distribution for θ, *i.e.* the density of the conditional random variable $\theta | Y_1 = 1$, as well as the Bayesian estimator for θ.

b. **[10pt]** We then ask a second student and observe $Y_2 = 0$. Recompute the posterior distribution for θ and the Bayesian estimator given this **additional** observation.

c. **[10pt]** After asking the 50 students enrolled in the class, it appears that only 10 of them would like to visit Paris. Compute the posterior distribution for θ given this information, as well as the Bayesian estimator.

d. **[10pt]** We want to compare the previous results with the maximum likelihood estimator. For this, we consider the following log-likelihood function:

$$\mathcal{L}(\theta) := \ln\big(\mathbb{P}(Y = y|\theta)\big),\ \theta \in [0, 1].$$

for any realisation y of a random variable Y. Compute the maximum likelihood estimator corresponding to the observations in questions **a.**, **b.** and **c.**. Briefly comment the results.

Recall. The Beta distribution with parameters $a, b > 0$ is characterised by a p.d.f. (density) given by

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

where Γ is the gamma function, which satisfies in particular $\Gamma(n) = (n-1)!$ for all $n \in \mathbb{N}^\star$. Moreover, the expectation of a random variable with Beta distribution (a, b) is equal to $a/(a+b)$.