ORF 418 - Optimal Learning
Emma Hubert

Midterm – Solution

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. **[5pt]** Compute the limiting reward $J(a^\star)$ for both $a^\star = 1$ and $a^\star = 2$.

Solution. Let $a^\star = 1$, meaning that we will only use the first arm for the exploitation part. Therefore, the distribution of the reward r_t is uniform on $[0, 1]$. By the law of large number, $J(a^\star)$ corresponds to the expected value of the rewards from the first arm, *i.e.*

$$J(a^\star) = \mathbb{E}\big[X\big] = 1/2.$$

Similarly, for $a^\star = 2$, the distribution of the reward r_t is Bernoulli with probability of success $p = 1/3$. By the law of large number, $J(a^\star)$ corresponds to the expected value of the rewards from the second arm, *i.e.*

$$J(a^\star) = \mathbb{E}\big[Y\big] = 1/3.$$

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

Solution. Let X the reward given by arm 1, with uniform distribution on $[0, 1]$, and let Y the reward given by arm 2, which distribution is a Bernoulli with probability of success $p = 1/3$. Now, $a^\star = 1$ if and only if $X > Y$, and $a^\star = 2$ iff $Y > X$ (note that $\mathbb{P}(X = Y) = 0$). To obtain the distribution of $a^\star$, we thus have to compute $\mathbb{P}(X > Y)$ and $\mathbb{P}(X < Y)$. We have:

$$\begin{aligned}\mathbb{P}(X > Y) &= \mathbb{P}(X > Y|Y = 0)\mathbb{P}(Y = 0) + \mathbb{P}(X > Y|Y = 1)\mathbb{P}(Y = 1)\\ &= 1 \times \mathbb{P}(Y = 0) + 0 \times \mathbb{P}(Y = 1) = \mathbb{P}(Y = 0) = 2/3.\end{aligned}$$

Therefore, the distribution of $a^\star$ is given by

$$a^\star = \begin{cases} 1 & \textbf{with probability } 2/3 \\ 2 & \textbf{with probability } 1/3. \end{cases}$$

We then use this to compute $\mathbb{E}[J(a^\star)]$:

$$\mathbb{E}[J(a^\star)] = \mathbb{E}[J(1)|a^\star = 1]\mathbb{P}(a^\star = 1) + \mathbb{E}[J(2)|a^\star = 2]\mathbb{P}(a^\star = 2) = \frac{1}{2} \times \frac{2}{3} + \frac{1}{3} \times \frac{1}{3} = \frac{4}{9}.$$

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

Solution. Let $X_1, X_2 \sim \mathcal{U}([0,1])$ be the (i.i.d) rewards given by arm 1, and let $Y_1, Y_2 \sim \mathcal{B}\text{er}(1/3)$ be the (i.i.d) rewards given by arm 2. Now, $a^\star = 1$ if and only if $\widehat{q}(1) > \widehat{q}(2)$, and $a^\star = 2$ iff $\widehat{q}(2) > \widehat{q}(1)$. To obtain the distribution of $a^\star$, we thus have to compute $\mathbb{P}(\widehat{q}(1) > \widehat{q}(2))$. We first remark that:

$$\mathbb{P}(\widehat{q}(1) > \widehat{q}(2)) = \mathbb{P}\bigg(\frac{1}{N}(X_1 + X_2) > \frac{1}{N}(Y_1 + Y_2)\bigg) = \mathbb{P}(X_1 + X_2 > Y_1 + Y_2).$$

First, since Y_1 and Y_2 are i.i.d. Bernoulli with probability of success $1/3$, the r.v. $\widetilde{Y} := Y_1 + Y_2$ has the following distribution:

$$\widetilde{Y} = \begin{cases} 0 & \textbf{with probability } 4/9 \\ 1 & \textbf{with probability } 4/9 \\ 2 & \textbf{with probability } 1/9. \end{cases}$$

Moreover, since X_1 and X_2 are i.i.d. uniform on $[0,1]$, the r.v. $\widetilde{X} := X_1 + X_2$ is uniformly distributed on $[0,2]$. We thus have that

$$\begin{aligned}\mathbb{P}(\widehat{q}(1) > \widehat{q}(2)) &= \mathbb{P}(\widetilde{X} > \widetilde{Y}) \\ &= \mathbb{P}(\widetilde{X} > \widetilde{Y} | \widetilde{Y} = 0)\mathbb{P}(\widetilde{Y} = 0) + \mathbb{P}(\widetilde{X} > \widetilde{Y} | \widetilde{Y} = 1)\mathbb{P}(\widetilde{Y} = 1) + \mathbb{P}(\widetilde{X} > \widetilde{Y} | \widetilde{Y} = 2)\mathbb{P}(\widetilde{Y} = 2) \\ &= 1 \times \frac{4}{9} + \frac{1}{2} \times \frac{4}{9} + 0 \times \frac{1}{9} = \frac{2}{3}.\end{aligned}$$

Therefore, the distribution of $a^\star$ is the same as before and we still have $\mathbb{E}[J(a^\star)] = 4/9$.

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?

Solution. In dimension one, we can easily compute that the controllability matrix $\mathcal{C} := (b)$ has full rank if and only if $b \neq 0$. Therefore, the system is controllable if and only if $b \neq 0$.

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?

Solution. The value function for this infinite horizon deterministic LQ control problem is defined as usual by:

$$v(x_0) := \inf_u J(x_0, u),\ x_0 \in \mathbb{R}.$$

For $b \neq 0$, we proved in the previous question that the system is controllable. Therefore, if $b \neq 0$, the value function is finite. It remains to study the case when $b = 0$. In this case, the dynamic of the state is given by

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

for some $a \in \mathbb{R}$ and $x_0 \in \mathbb{R}$ arbitrary fixed. By a straightforward mathematical induction, one can prove that $x_k = a^k x_0$ for all $k \geq 0$. In particular, the state x is not controlled (because $b = 0$). We can thus rewrite the value function as follows:

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

The infimum in the previous equation is clearly equal to 0 (take $u_k = 0$ for all $k \geq 0$). We thus have:

$$v(x_0) = 2x_0^2 \sum_{k=0}^{\infty} \frac{1}{2^k} a^{2k} = 2x_0^2 \sum_{k=0}^{\infty} \left(\frac{a^2}{2}\right)^k.$$

If $x_0 = 0$, then $v(x_0) = 0$ which is clearly finite. If $x_0 \neq 0$, then we need to have $|a^2/2| < 1$, *i.e.* $|a| < \sqrt{2}$, to ensure convergence of the sum. To summarise, the value function is infinite if and only if $b = 0$, $x_0 \neq 0$, and $|a| > \sqrt{2}$.

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:

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

Solution. The general Dynamic Programming equation (Thm 2.2.1 of the Lecture notes) is:

$$v(x) = x^\top M x + \inf_{u \in \mathbb{R}^\ell} \left\{ u^\top N u + \rho v(Ax + Bu) \right\}, \quad \textbf{for all} \quad x \in \mathbb{R}^d.$$

Here, we are in dimension 1 ($d = \ell = 1$) with parameters $\rho = 1/2$, $A = a \in \mathbb{R}$, $B = b \neq 0$, $M = 2$ and $N = 3$. Therefore, the DP equation becomes:

$$v(x) = 2x^2 + \inf_{u \in \mathbb{R}} \left\{ 3u^2 + \frac{1}{2} v(ax + bu) \right\}, \quad \textbf{for all} \quad x \in \mathbb{R}.$$

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 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):

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

Solution. Using the assumption $v(x) = Vx^2$ for all $x \in \mathbb{R}$ in (0.1), we obtain:

$$Vx^2 = 2x^2 + \inf_{u \in \mathbb{R}} f(x, u), \quad \textbf{with} \quad f(x, u) := 3u^2 + \frac{1}{2} V(ax + bu)^2.$$

To compute $\inf_{u \in \mathbb{R}} f(x, u)$, we compute the FOC (with respect to u):

$$\partial_u f(x, u) = 6u + bV(ax + bu) = u(6 + Vb^2) + abVx = 0,$$

and the optimal is therefore $u^\star(x) = \dfrac{-abV}{6 + Vb^2} x$. Note that the SOC, *i.e.* $6 + Vb^2$ is satisfied since V should be positive. We now compute:

$$\begin{aligned} \inf_{u \in \mathbb{R}} f(x, u) = f(x, u^\star) &= 3\left(\frac{-abV}{6 + Vb^2} x\right)^2 + \frac{1}{2} V \left(ax - \frac{ab^2V}{6 + Vb^2} x\right)^2 = \left(\frac{3a^2b^2V^2}{(6 + Vb^2)^2} + \frac{18Va^2}{(6 + Vb^2)^2}\right) x^2 \\ &= \frac{3Va^2x^2(b^2V + 6)}{(6 + Vb^2)^2} = \frac{3Va^2x^2}{6 + Vb^2}. \end{aligned}$$

Using this in the DP equation, we obtain:

$$Vx^2 = 2x^2 + \frac{3Va^2x^2}{6+Vb^2}, \qquad \textbf{for all } \ x \in \mathbb{R},$$

which is equivalent to

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

The following is not required but presented here for the sake of completeness. **The discriminant Δ of this quadratic equation is positive since**

$$\Delta := (6 - 2b^2 - 3a^2)^2 + 48b^2 > 0.$$

The equation therefore has two solutions (recall that $b \neq 0$):

$$V_1 = \frac{-(6 - 2b^2 - 3a^2) + \sqrt{\Delta}}{2b^2}, \ \textbf{ and } \ V_2 = \frac{-(6 - 2b^2 - 3a^2) - \sqrt{\Delta}}{2b^2}$$

One can verify that V_1 is positive, since

$$\sqrt{\Delta} = \sqrt{(6 - 2b^2 - 3a^2)^2 + 48b^2} \geq \sqrt{(6 - 2b^2 - 3a^2)^2} = |6 - 2b^2 - 3a^2|,$$

and is therefore the appropriate solution.

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.

Solution. To compute the density of $\theta|Y_1 = 1$, we use the following Bayes formula:

$$f_{\theta|Y_1=1}(x) = \frac{\mathbb{P}(Y_1 = 1|\theta = x)f_\theta(x)}{\mathbb{P}(Y_1 = 1)}, \ x \in (0, 1).$$

Here we have:

$$f_\theta(x) := \mathbb{1}_{x\in(0,1)}, \ \mathbb{P}(Y_1 = 1|\theta = x) = x, \ \textbf{ and } \ \mathbb{P}(Y_1 = 1) = \int_0^1 \mathbb{P}(Y_1 = 1|\theta = x)f_\theta(x)\mathrm{d}x = \int_0^1 x\mathrm{d}x = \frac{1}{2}.$$

We can thus compute the posterior density:

$$f_{\theta|Y_1=1}(x) = 2x\mathbb{1}_{x\in(0,1)}.$$

The Bayesian estimator for θ is given by

$$\widehat{\theta} = \mathbb{E}[\theta|Y_1 = 1] = \int_{\mathbb{R}} x f_{\theta|Y_1=1}(x)\mathrm{d}x = 2\int_0^1 x^2\mathrm{d}x = \frac{2}{3}.$$

b. **[10pt]** We then ask a second student and observe $Y_2 = 0$. Recompute the posterior distribution for θ and the Bayesian estimator given the observation of $Y_1 = 1$ **and** $Y_2 = 0$.

Solution. Recall that Y_1 and Y_2 are i.i.d Bernoulli with success probability θ. To compute the density of $\theta|Y_1 = 1, Y_2 = 0$, we use the following Bayes formula:

$$f_{\theta|Y_1=1,Y_2=0}(x) = \frac{\mathbb{P}(Y_1 = 1, Y_2 = 0|\theta = x)f_\theta(x)}{\mathbb{P}(Y_1 = 1, Y_2 = 0)}, \ x \in (0, 1).$$

Here we have:

$$f_\theta(x) := \mathbb{1}_{x\in(0,1)},\ \mathbb{P}(Y_1=1,Y_2=0|\theta=x) = x(1-x),$$
$$\textbf{and}\ \ \mathbb{P}(Y_1=1,Y_2=0) = \int_0^1 \mathbb{P}(Y_1=1,Y_2=0|\theta=x) f_\theta(x)\mathrm{d}x = \int_0^1 x(1-x)\mathrm{d}x = \frac{1}{6}.$$

We can thus compute the posterior density:

$$f_{\theta|Y_1=1,Y_2=0}(x) = 6x(1-x)\mathbb{1}_{x\in(0,1)}.$$

We can compute the expectation "by hand" as before, or recognize a Beta distribution with $a = b = 2$, since

$$\frac{\Gamma(4)}{\Gamma(2)\Gamma(2)} = 6$$

Therefore, the Bayesian estimator for θ is given by

$$\widehat{\theta} = \mathbb{E}[\theta|Y_1=1,Y_2=0] = \frac{a}{a+b} = \frac{1}{2}.$$

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 for θ.

Solution. Let $Y = Y_1 + \ldots + Y_2$. Recall that Y_k are i.i.d Bernoulli with success probability θ. Therefore, Y is a Binomial $(50,\theta)$. To compute the density of $\theta|Y=10$, we use the same Bayes formula as question a), but here with:

$$f_\theta(x) := \mathbb{1}_{x\in(0,1)},\ \mathbb{P}(Y=10|\theta=x) = \binom{50}{10}x^{10}(1-x)^{40},\ \ \textbf{and}\ \ \mathbb{P}(Y=10) = \binom{50}{10}\int_0^1 x^{10}(1-x)^{40}\mathrm{d}x.$$

We thus obtain:

$$f_{\theta|Y=10}(x) = \frac{1}{C}x^{10}(1-x)^{40}\mathbb{1}_{x\in(0,1)},\ \ \textbf{with}\ \ C := \int_0^1 x^{10}(1-x)^{40}\mathrm{d}x.$$

Using the fact that the integral of any density is equal to 1, we have:

$$C := \int_0^1 x^{10}(1-x)^{40}\mathrm{d}x = \frac{\Gamma(11)\Gamma(41)}{\Gamma(52)}.$$

We thus recognize that the density

$$f_{\theta|Y=10}(x) = \frac{\Gamma(52)}{\Gamma(11)\Gamma(41)}x^{10}(1-x)^{40}\mathbb{1}_{x\in(0,1)},$$

is the density of a Beta distribution with $a = 11$ and $b = 41$. Therefore, the Bayesian estimator for θ is given by

$$\widehat{\theta} = \mathbb{E}[\theta|Y=10] = \frac{a}{a+b} = \frac{11}{52}.$$

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. and b.

Solution. For question a., we have the observation $Y = 1$, for Y a Bernoulli with probability of success θ. The log-likelihood function is thus given by

$$\mathcal{L}(\theta) = \ln\big(\mathbb{P}(Y = 1|\theta)\big) = \ln(\theta).$$

As the $\ln$ function is increasing, this is maximised for $\theta = 1 > \widehat{\theta} = 2/3$.

For question b., we have the observation $Y := (Y_1, Y_2) = (1, 0)$, for (Y_1, Y_2) i.i.d Bernoulli with probability of success θ. The log-likelihood function is thus given by

$$\mathcal{L}(\theta) = \ln\big(\mathbb{P}(Y_1 = 1, Y_2 = 0|\theta)\big) = \ln\big(\theta(1-\theta)\big) = \ln(\theta) + \ln(1-\theta).$$

The FOC condition gives

$$\frac{1}{\theta} - \frac{1}{1-\theta} = 0, \qquad \textbf{\textit{i.e.}} \qquad \theta = \frac{1}{2} = \widehat{\theta}.$$

Finally, for question c., we have the observation $Y := Y_1 + \ldots + Y_{50} = 10$, for Y Binomial with $n = 50$ and probability of success θ. The log-likelihood function is thus given by

$$\mathcal{L}(\theta) = \ln\big(\mathbb{P}(Y = 10|\theta)\big) = \ln\left(\binom{50}{10}\theta^{10}(1-\theta)^{40}\right) = \ln\left(\binom{50}{10}\right) + 10\ln(\theta) + 40\ln(1-\theta).$$

The FOC condition gives

$$\frac{10}{\theta} - \frac{40}{1-\theta} = 0, \qquad \textbf{\textit{i.e.}} \qquad \theta = \frac{1}{5} < \widehat{\theta}.$$

Hint. 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 $\Gamma(1) = 1$ and $\Gamma(x+1) = x\Gamma(x)$. Moreover, its mean is equal to $a/(a+b)$.