Exam of the course *Markov decision processes : dynamic programming and applications*

Marianne Akian

ENSTA Course SOD312 & M2 Optimization (Paris-Saclay University and IP Paris)

Exam of Mardi 12 janvier 2021 (Durée 3h)
Revised version with solution

This text contains 2 different exams :

M2 Exam consists in Problems 1 and 3 it is for the students who need to validate the full lectures (to obtain a M2 mark or to obtain more ECTS). No mark will be given to answers to questions of Problem 2 for these students.

ENSTA Exam consists in Problems 1 and 2 it is for the other students (ENSTA students that only need to validate the ENSTA lectures). No mark will be given to answers to questions of Problem 3 for these students (moreover this problem may use notions that were not teached to ENSTA students).

The problems are independent and often questions can be solved without solving the previous questions. The solution can be written either in French or English. *Documents (handwritten or typed courses and exercises notes, together with books related to the course) are allowed.*

1 Problem 1 (for all students)

An investor would like to buy a flat in Paris. To simplify the study, we assume that each day there is exactly one advertisement for one flat, that the investor is the first person to visit it and that he decides immediatly to buy it or not. We denote by a_k the surface area of the flat advertised on the kth day, by m_k its price per square meter, and by $p_k = m_k a_k$ its price. The investor has enough money to buy any flat at any price, but wishes to maximize the surface area and to minimize the price of the unique flat he will buy.

We assume here that $(a_k)_{k\geq 0}$ is a sequence of independent identically distributed random variables, that $(m_k)_{k\geq 0}$ is a stationary Markov chain, that the two sequences are independent, and that they both take values in finite subsets $\mathcal{A}$ and $\mathcal{M}$ of $(0, +\infty)$ respectively. We denote by q the law of the a_k : $P(a_k = a) = q(a)$ and by M the transition matrix of the Markov chain $(m_k)_{k\geq 0}$.

Q 1.1. Show that the sequence $X_k := (a_k, m_k)$ is a Markov chain over the finite set $\mathcal{A} \times \mathcal{M}$ and determine its transition probabilities.

Q 1.2. We denote by τ the day in which the investor is buying a flat. Equivalently, we denote by U_k a random variable equal to 1 when the investor is bying a flat the kth day or before and 0 otherwise, and consider the process, for $k \geq 0$, $Y_k = f(a_k, m_k, U_{k-1})$ in which $f(a, m, 0) = (a, m)$ and $f(a, m, 1) = c$, where c is an element not in $\mathcal{A} \times \mathcal{M}$, and by convention $U_{-1} = 0$.

We denote by $\mathcal{E}$ the disjoint union of $\mathcal{A} \times \mathcal{M}$ and $\{c\}$. Show that the sequence $(Y_k, U_k)_{k\geq 0}$ can be seen as a Markov Decision Process and determine its action spaces $\mathcal{C}(y)$ and its transition probabilities $M_{yy'}^{(u)}$ when $y, y' \in \mathcal{E}$ and $u \in \mathcal{C}(y)$.

Q 1.3. Let us consider the utility function $G : (0, +\infty)^2 \to (0, +\infty)$ such that $G(a, p) = a^\gamma p^{-\gamma'}$, where $0 < \gamma < \gamma'$ (hence G is nondecreasing with respect to a and nonincreasing with respect to p). Assume that the investor is stopping looking for a flat after T days, whether he bought one or not. Assume also that his choice depends only on his past actions and on the characteristics of the flats he has already visited, and that he want to maximize the expected utility of the flat he is buying. Show that this is equivalent to maximize

$$\mathbb{E}\left[\sum_{k=0}^{T-1} R(Y_k, U_k)\right] \ ,$$

over all strategies with respect to the above MDP, where days are numbered from 0, and for all $u \in \{0, 1\}$,

$$R(y, u) = \begin{cases} G(a, ma)u & \text{when } y = (a, m) \in \mathcal{A} \times \mathcal{M} \\ 0 & \text{when } y = c \ . \end{cases}$$

Show that this is equivalent to the maximization of

$$\mathbb{E}\left[G_\tau(a_\tau, m_\tau a_\tau)\right]$$

over all stopping times $\tau \leq T$ with respect to the Markov chain $(X_k)_{k\geq 0}$, with $G_T = 0$ and $G_k = G$ when $k \leq T - 1$.

Q 1.4. Let $(v_n)_{0\leq n\leq T}$ be the value function of the above Markov decision problem (or stopping time problem) :

$$v_n(y) := \max \mathbb{E}\left[\sum_{k=n}^{T-1} R(Y_k, U_k) \mid Y_n = y\right] \ ,$$

where the maximization holds over all strategies associated to the above MDP. Write a recurrence equation for v_n, and explain how the investor can choose the sequence U_n or the stopping time τ, by using this equation.

Q 1.5. Consider a variant of the problem in which the investor may take two days to decide if he buy a flat : he want to make an offer only for the best flat among the one of the current day and of the day before. We also assume that there is no chance that the flat of the day before is sold out when the investor takes his decision. Then, this problem is equivalent to the maximization of

$$\mathbb{E}\left[\max(G_\tau(a_\tau, m_\tau a_\tau), G_{\tau-1}(a_{\tau-1}, m_{\tau-1} a_{\tau-1})\right]$$

over all stopping times $\tau \leq T$ with respect to the Markov chain $(X_k)_{k\geq 0}$.

Can you write this problem as a Markov Decision Problem ?

Q 1.6. Can you do the same when there is a probability $1/2$ that the flat of the day before is sold out, and the investor will do an offer for the best available flat among the one of the current day and the one of the day before (if it is available) ?

2 Problem 2 (to validate the ENSTA lectures only)

We consider a variant of Problem 1. The notations $a_k, m_k, p_k = m_k a_k$ are the same. The sequence $(a_k)_{k\geq 0}$ is again a sequence of independent identically distributed random variables. However, we assume now that $m_k = \delta_k m_{k-1}$, where δ_k is a growth factor and $(\delta_k)_{k\geq 0}$ is a stationary Markov chain. The sequences $(a_k)_{k\geq 0}$ and $(\delta_k)_{k\geq 0}$ are independent, and they both take values in finite subsets $\mathcal{A}$ and $\mathcal{D}$ of $(0, +\infty)$ respectively. We still denote by q the law of the a_k : $P(a_k = a) = q(a)$, but now M is the transition matrix of the Markov chain $(\delta_k)_{k\geq 0}$.

Q 2.1. Show that the sequence $X_k := (a_k, \delta_k)$ is a Markov chain over the finite set $\mathcal{A} \times \mathcal{D}$ and determine its transition probabilities.

Q 2.2. Is (a_k, m_k) or (a_k, δ_k, m_k) a Markov chain over a finite set ?

Q 2.3. We use the same notations τ and U_k as in Problem 1 : τ is the day in which the investor is buying a flat and U_k is a random variable equal to 1 when the investor is bying a flat the kth day or before and 0 otherwise.

We consider now the process : $Z_k = f(a_k, \delta_k, U_{k-1})$ in which $f(a, \delta, 0) = (a, \delta)$ and $f(a, \delta, 1) = c$, where c is an element not in $\mathcal{A} \times \mathcal{D}$. We also denote by $\mathcal{E}'$ the disjoint union of $\mathcal{A} \times \mathcal{D}$ and $\{c\}$. Similarly to Problem 1, the sequence $(Z_k, U_k)_{k\geq 0}$ can be seen as a Markov Decision Process over the state space $\mathcal{E}'$.

We consider the same utility function $G : (0, +\infty)^2 \to (0, +\infty)$ such that $G(a, p) = a^\gamma p^{-\gamma'}$, where $0 < \gamma' < \gamma$, and assume (as in Problem 1) that the investor want to maximize the expected utility of the flat he is buying :

$$\mathbb{E}\left[G_\tau(a_\tau, m_\tau a_\tau)\right] \quad .$$

Show that this problem can be seen (up to a positive multiplicative factor) as a problem associated to the Markov Decision Process $(Z, U) := (Z_k, U_k)_{k\geq 0}$, with the following finite horizon mixed criteria :

$$J(Z, U) := \mathbb{E}\left[\sum_{k=0}^{T-1} \left(\prod_{\ell=0}^{k-1} \alpha(Z_\ell, U_\ell)\right) r(Z_k, U_k)\right], \tag{1}$$

where the instantaneous reward r and variable discount factor α are defined, for all $u \in \{0, 1\}$, by

$$r(z, u) = \begin{cases} \delta^{-\gamma'} a^{\gamma-\gamma'} u & \text{when } z = (a, \delta) \in \mathcal{A} \times \mathcal{D} \\ 0 & \text{when } z = c \ . \end{cases}$$

and

$$\alpha(z, u) = \begin{cases} \delta^{-\gamma'} & \text{when } z = (a, \delta) \in \mathcal{A} \times \mathcal{D} \\ 0 & \text{when } z = c \ . \end{cases}$$

Q 2.4. Let $v^T(z)$ be the value of this problem when the initial state is equal to $z \in \mathcal{E}'$, and when the horizon is equal to T :

$$v^T(z) = \max \mathbb{E}\left[\sum_{k=0}^{T-1} \left(\prod_{\ell=0}^{k-1} \alpha(Z_\ell, U_\ell)\right) r(Z_k, U_k) \mid Z_0 = z\right] \quad ,$$

where the maximum is taken over all (relaxed) strategies. Show that v^T satisfies the following recurrence equation :

$$v^T(z) = \delta^{-\gamma'} \max(\mathbb{E}\left[v^{T-1}(a_1, \delta_1) \mid \delta_0 = \delta\right], a^{\gamma-\gamma'}), \quad \forall z = (a, \delta) \in \mathcal{A} \times \mathcal{D} \ , \tag{2}$$

with the initial condition $v^0 = 0$.

Q 2.5. For all $k \geq 0$, let $\pi^k \in \{0,1\}^{\mathcal{E}'}$ be such that $\pi^k(z) = 1$ if $z = c$ or if $z = (a, \delta) \in \mathcal{A} \times \mathcal{D}$ is such that the maximum in (2) is attained by the second term, and $\pi^k(z) = 0$ otherwise. Construct an optimal strategy $\sigma = (\sigma_k)_{0 \leq k \leq T}$ for the above Markov Decision Problem by using the maps π^k, $k \geq 0$. Deduce also the corresponding optimal stopping time.

Q 2.6. Deduce that there exists an optimal policy of threshold type : $\pi^T(z) = 1$ when $z = (a, \delta) \in \mathcal{A} \times \mathcal{D}$ is such that $a \geq \underline{a}^T(\delta)$ for some map $\underline{a}^T : \mathcal{D} \to \mathcal{A} \cup \{+\infty\}$ (where $+\infty$ means that $\pi^T(z) = 1$ never holds).

Q 2.7. Let $\mathcal{B}$ be the Bellman operator from $\mathbb{R}^{\mathcal{A} \times \mathcal{D}}$ corresponding to (2) :

$$[\mathcal{B}(v)](a, \delta) = \delta^{-\gamma'} \max(\mathbb{E}\left[v(a_1, \delta_1) \mid \delta_0 = \delta\right], a^{\gamma - \gamma'}), \quad \forall z = (a, \delta) \in \mathcal{A} \times \mathcal{D} \ .$$

Show, with the help of $\mathcal{B}$, that v^T is nondecreasing with respect to T : $v^T \leq v^{T+1}$. Deduce that $\underline{a}^T$ is also nondecreasing with respect to T.

Q 2.8. Assume that there exists $\beta \leq 1$ such that $\mathbb{E}\left[\delta_1^{-\gamma'} \mid \delta_0 = \delta\right] \leq \beta$ for all $\delta \in \mathcal{D}$. Show that the sequences $(v^T)_{T \geq 0}$ is bounded from above (by some constant function for instance).

Q 2.9. Deduce that v^T and $\underline{a}^T$ converge towards some maps v^∞ and $\underline{a}^\infty$, when T goes to infinity. What are the interpretations of these maps ? Which equations do they satisfy ?

Q 2.10. Assume that $\beta < 1$. Show that $\mathcal{B}$ is contracting with constant β for the following norm on $\mathbb{R}^{\mathcal{A} \times \mathcal{D}}$:

$$\|v\| := \max_{(a,\delta) \in \mathcal{A} \times \mathcal{D}} |\delta^{\gamma'} v(a, \delta)| \ .$$

What this means for the equations of Question 2.9 ?

Q 2.11. Denote $w(a, \delta) = \delta^{\gamma'} v^\infty(a, \delta)$. Write an equation for w. Interpret it as the stationnary dynamic programming equation of a discounted infinite horizon or stopping time problem for a new Markov Decision Process on $\mathcal{E}'$ or $\mathcal{A} \times \mathcal{D}$ to be determined.

Q 2.12. What is the policy iteration algorithm computing w (and thus v^∞) ? How many steps are needed for such an algorithm in general ?

Q 2.13. Assume now that δ_k is a sequence of independent random variables with same law, and let $\beta = \mathbb{E}\left[\delta_0^{-\gamma'}\right] < 1$. Show that, in that case, w, v^∞ and $\underline{a}^\infty$ can be computed easily as functions of $\bar{v} = \mathbb{E}\left[v^\infty(a_0, \delta_0)\right]$. Show also that $\bar{v}$ is the fixed point of an operator $\overline{\mathcal{B}} : \mathbb{R} \to \mathbb{R}$ ($\bar{v} = \overline{\mathcal{B}}(\bar{v})$) which is monotone, contracting, convex, and piecewise affine.

Q 2.14. Give an upper bound N on the number of regions in which $\overline{\mathcal{B}}$ is affine. Interpret $\overline{\mathcal{B}}$ as the dynamic programming operator of a discounted infinite horizon problem with a singleton state space and at most N actions. Deduce a bound on the number of Policy Iterations for this equation. Can we bound also the number of Policy Iterations of Question 2.12 in that case ?

3 Problem 3 (to validate the full M2 lectures only)

Let $\mathcal{E}$ and $\mathcal{C}$ be finite sets and for all $x \in \mathcal{E}$, let $\mathcal{C}(x)$ be a subset of $\mathcal{C}$. Let us consider the operator $\mathcal{B} : \mathbb{R}^{\mathcal{E}} \to \mathbb{R}^{\mathcal{E}}$:

$$[\mathcal{B}(v)](x) = \max_{u \in \mathcal{C}(x)} \left(r(x,u) + \sum_{y \in \mathcal{E}} M_{xy}^{(u)} v(y) \right), \quad x \in \mathcal{E} \ ,$$

where $r : \mathcal{E} \times \mathcal{C} \to \mathbb{R}$ and for all $x \in \mathcal{E}$ and $u \in \mathcal{C}(x)$, $(M_{xy}^{(u)})_{y \in \mathcal{E}}$ is a probability vector on $\mathcal{E}$.

Q 3.1. Explain for which Markov Decision Processes $\mathcal{B}$ is the dynamic programming operator (explain the parameters). What is the meaning of a solution $\rho \in \mathbb{R}$ and $v \in \mathbb{R}^{\mathcal{E}}$ to the following equation

$$\rho \mathbf{1} + v = \mathcal{B}(v) \ ?$$

Q 3.2. For all $v \in \mathbb{R}^{\mathcal{E}}$, we denote $\mathsf{t}(v) := \max_{x \in \mathcal{E}} v(x)$ and $\mathsf{b}(v) := \min_{x \in \mathcal{E}} v(x)$. Show that

$$\mathsf{t}(\mathcal{B}(v) - \mathcal{B}(w)) \leq \mathsf{t}(v - w), \quad \text{and} \quad \mathsf{b}(\mathcal{B}(v) - \mathcal{B}(w)) \geq \mathsf{b}(v - w) \ .$$

Q 3.3. Let $0 \leq \alpha < 1$. Show that the operator $v \in \mathbb{R}^{\mathcal{E}} \mapsto \mathcal{B}(\alpha v) \in \mathbb{R}^{\mathcal{E}}$ has a unique fixed point, that we shall denote by $\xi_\alpha(\mathcal{B})$. What is the meaning of $\xi_\alpha(\mathcal{B})$ in terms of Markov Decision Processes ?

Q 3.4. Show, using the previous questions, that

$$\mathsf{t}(\xi_\alpha(\mathcal{B})) \leq \frac{\mathsf{t}(\mathcal{B}(0))}{1-\alpha}, \quad \text{and} \quad \mathsf{b}(\xi_\alpha(\mathcal{B})) \geq \frac{\mathsf{b}(\mathcal{B}(0))}{1-\alpha},$$

and deduce that $(1-\alpha)\xi_\alpha(\mathcal{B})$ is bounded when $\alpha \to 1^-$.

Denote by $\Pi = \{\pi : \mathcal{E} \to \mathcal{C} \mid \pi(x) \in \mathcal{C}(x) \ \forall \in \mathcal{E}\}$ the set of (stationary) policies for the Markov decision problem associated to the operator $\mathcal{B}$, and for each $\pi \in \Pi$, denote by $r^{(\pi)} \in \mathbb{R}^{\mathcal{E}}$ the vector with entries $r_x^{(\pi)} = r(x, \pi(x))$, by $M^{(\pi)} \in \mathbb{R}^{\mathcal{E} \times \mathcal{E}}$ the Markov matrix with entries $M_{xy}^{(\pi)} = M_{xy}^{(\pi(x))}$, and by $\mathcal{B}^{(\pi)}$ the affine operator :

$$\mathcal{B}^{(\pi)}(v) = r^{(\pi)} + M^{(\pi)} v \ .$$

Recall that, since the sets $\mathcal{E}$ and $\mathcal{C}$ are finite, we have

$$\mathcal{B}(v) = \max_{\pi \in \Pi} \mathcal{B}^{(\pi)}(v), \quad \forall v \in \mathbb{R}^{\mathcal{E}} \ .$$

Q 3.5. Interpret $\mathcal{B}^{(\pi)}$ as a dynamic programming operator, and explain why

$$\xi_\alpha(\mathcal{B}) = \max_{\pi \in \Pi} \xi_\alpha(\mathcal{B}^{(\pi)}).$$

Q 3.6. Recall (from the course) that for all $\pi \in \Pi$, $\xi_\alpha(\mathcal{B}^{(\pi)})$ has an asymptotics expansion of the following form, when $\alpha \to 1^-$:

$$\xi_\alpha(\mathcal{B}^{(\pi)}) = \frac{v_{-1}}{1-\alpha} + v_0 + (1-\alpha)v_1 + (1-\alpha)^2 v_2 + \cdots$$

where $v_{-1}, v_0, v_1, v_2, \ldots \in \mathbb{R}^{\mathcal{E}}$ depend on the policy π. Recall also that there exists a Blackwell optimal policy $\pi \in \Pi$, that is there exists $0 \leq \alpha_0 < 1$ such that

$$\xi_\alpha(\mathcal{B}) = \xi_\alpha(\mathcal{B}^{(\pi)}), \qquad \forall \alpha \in [\alpha_0, 1) \ .$$

Show that the vectors v_{-1} and v_0 associated to the Blackwell optimal policy satisfy

$$M^{(\pi)}v_{-1} = v_{-1}, \qquad v_0 + v_{-1} = r^{(\pi)} + M^{(\pi)}v_0, \qquad P^{(\pi)}v_0 = 0 \ ,$$

where $P^{(\pi)}$ denotes the spectral projector of $M^{(\pi)}$ for the eigenvalue 1.

Show that the pair $(v_1, v_0) \in (\mathbb{R}^{\mathcal{E}})^2$ solution of the above equations is unique.

Q 3.7. Deduce that there exist $v, \eta \in \mathbb{R}^{\mathcal{E}}$ such that

$$\mathcal{B}^{(\pi)}(v + t\eta) = v + (t+1)\eta \ , \forall t \geq 0 \ .$$

Q 3.8. Show that the same holds for $\mathcal{B}$ (use that $\mathcal{B}$ is piecewise affine).

Q 3.9. Show that there exists $t_0 \geq 0$ such $\mathcal{B}(v+t\eta) = v+(t+1)\eta \ , \forall t \geq t_0$ if and only if (v, η) satisfies :

$$\eta(x) = \max_{u \in \mathcal{C}(x)} \left(\sum_{y \in \mathcal{E}} M^{(u)}_{xy} \eta(y) \right), \quad \forall x \in \mathcal{E} \ , \tag{3a}$$

$$\eta(x) + v(x) = \max_{u \in S(x)} \left(r(x,u) + \sum_{y \in \mathcal{E}} M^{(u)}_{xy} v(y) \right) \quad \text{with } S(x) := \operatorname*{Argmax}_{u \in \mathcal{C}(x)} \left(\sum_{y \in \mathcal{E}} M^{(u)}_{xy} \eta(y) \right), \quad \forall x \in \mathcal{E} \ . \tag{3b}$$

Q 3.10. Show that for all $\varphi \in \mathbb{R}^{\mathcal{E}}$, we have

$$\eta = \lim_{k \to \infty} \frac{1}{k} \mathcal{B}^k(\varphi) = \lim_{\alpha \to 1^-} (1-\alpha)\xi_\alpha(\mathcal{B}) \ .$$

What can we conclude on a solution (η, v) of (3) ?

Q 3.11. Deduce that for all $x \in \mathcal{E}$ and all (relaxed) strategies σ, we have

$$\eta(x) \geq J^{(+,\sigma)}(x) := \limsup_{T \to \infty} \left\{ \frac{1}{T} \mathbb{E} \left[\sum_{k=0}^{T} r(X_k, U_k) \mid X_0 = x \right] \right\} \ ,$$

where (X_k, U_k) is the process induced by σ.

Q 3.12. Deduce that the value of the mean-payoff Markov Decision Problem with the above parameters is equal to η and that any Blackwell optimal policy is optimal for this problem.

Q 3.13. Consider the Linear Program :

$$\lambda_0 := \min\{\lambda \mid \lambda \in \mathbb{R}, \ v \in \mathbb{R}^{\mathcal{E}}, \ -v(x) + r(x,u) + \sum_{y \in \mathcal{E}} M^{(u)}_{xy} v(y) \leq \lambda \quad \forall x \in \mathcal{E}, \ u \in \mathcal{C}(x)\} \ .$$

Show (using the previous questions) that the value λ_0 satisfies :

$$\lambda_0 = \mathsf{t}(\eta) \ .$$

Q 3.14. Denote $\mathcal{A} := \{(x,u) \mid x \in \mathcal{E}, \ u \in \mathcal{C}(x)\}$. How the previous linear program is related to the following one ?

$$J_0 := \max\{ \sum_{(x,u) \in \mathcal{A}} (r(x,u) f(x,u)) \mid f \in \mathbb{R}^{\mathcal{A}}_+, \quad \text{satisfying (4)}\}$$

with :

$$\sum_{(x,u) \in \mathcal{A}} f(x,u) = 1, \quad \text{and} \quad \sum_{u' \in \mathcal{C}(y)} f(y,u') = \sum_{(x,u) \in \mathcal{A}} M^{(u)}_{xy} f(x,u), \quad \forall y \in \mathcal{E} \ . \tag{4}$$

Q 3.15. What happens when the graph of the Markov Decision Process is strongly connected ?