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

Marianne Akian

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

Mardi 8 novembre 2022
Durée 3h

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).

Problems 1,2 and 3 are independent. 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)

Consider a problem of conservation of a threatened species in some environmentally protected zone. Let X_n denotes the number of population of the threatened species during the nth year (years are numbered $0, 1, 2, ...$), and U_n be the number of population of predators of the threatened species during the same year. We assume that U_n can be choosen, and that there are positive integers $\ell \leq M$, such that $X_n \in \mathcal{E} = \{0, 1, \dots, M\}$, $U_n \in \mathcal{C} = \{0, 1, \dots, M\}$, and

$$X_{n+1} = \begin{cases} \lceil \alpha_{n+1} X_n (M + 1 - (U_n + X_n)) \rceil & \text{if } U_n + X_n \leq M \text{ and } X_n > \ell \\ 0 & \text{otherwise,} \end{cases}$$

where $\lceil x \rceil$ denotes the least integer that is greater than or equal to x, $(\alpha_n)_{n\geq 1}$ is a sequence of identically distributed independent random variables with positive values ($(M+1)\alpha_n$ represents the growth factor of the population when this population is small enough, α_n may depend on weather conditions but not on the number of the population itself).

Q 1.1. Explain why $(X_n)_{n\geq 0}$ is the sequence of states and $(U_n)_{n\geq 0}$ the sequence of controls of a stationary Markov decision process and give the transition probabilities $M_{xy}^{(u)}$. Explain what is the best choice of the constrained control sets $\mathcal{C}(x)$.

Q 1.2. In order to avoid extinction of the given threatened species, we choose appropriate constrained control sets $\mathcal{C}(x)$ and would like to find a (pure or relaxed) strategy minimizing the probability that $X_N \leq \ell$, for some given N. Write this problem as a Markov decision problem with finite horizon N.

Q 1.3. What is the Dynamic programming equation satisfied by the value function of this problem ? Explain how an optimal strategy can be obtained.

Q 1.4. We replace the previous criterion by considering the following problem :

$$\max \mathbb{P}\left(X_1 + \cdots + X_N \geq Nh \mid X_0 = x\right) \quad ,$$

where $h \in (\ell, 1)$, $(X_n)_{n\geq 0}$ is the sequence of states of the above MDP, and the maximization holds over the set of all (pure or relaxed) strategies. Write this problem as a Markov decision problem with finite horizon N, and enlarged state space.

Q 1.5. What are now the corresponding dynamic programming equation, and the optimal policies ?

Q 1.6. Since extinction is still possible (with positive probability), we would like now to maximize the expected extinction time if it occurs before the Nth year. Explain why this is equivalent to the following Markov decision problem (still with finite horizon N)

$$\max \mathbb{E}\left[\tau \mid X_0 = x\right] \quad ,$$

where τ is the first time $\leq N$ such that $X_\tau \leq \ell$, and the maximization holds over the set of all (pure or relaxed) strategies. Give the corresponding dynamic programming equation, and determine an optimal strategy.

---

2 Problem 2 (to validate the ENSTA lectures only)

Q 2.1. Consider two independent Markov chains $(Y_n^1)_{n\geq 0}$ and $(Y_n^2)_{n\geq 0}$ taking their values in the finite state spaces $\mathcal{E}^1$ and $\mathcal{E}^2$ respectively, with transition matrices M^1 and M^2. We built a Markov decision process with state space $\mathcal{E} = \mathcal{E}^1 \times \mathcal{E}^2$ and action space $\mathcal{C} = \{1, 2\}$, with the following transition matrix :

$$M^{(u)}_{(x_1,x_2),(x_1',x_2')} := \begin{cases} M^1_{x_1 x_1'} \delta_{x_2 x_2'} & \text{if } u = 1 \ , \\ M^2_{x_2 x_2'} \delta_{x_1 x_1'} & \text{if } u = 2 \ , \end{cases}$$

where $\delta_{xx'} = 1$ if $x = x'$ and $\delta_{xx'} = 0$ otherwise. Explain the relation between the coordinates X_n^1 and X_n^2 of a state sequence $X_n \in \mathcal{E}$ of the Markov decision process (associated to any strategy, for instance a pure stationary Markov strategy π, so that $U_n = \pi(X_n)$) and the Markov chains Y_n^1 and Y_n^2.

Q 2.2. Let the instantaneous reward at each time n of the process be equal to :

$$r(u, x) = r_i(x_i) \quad \text{for } u = i \in \{1, 2\}, \text{ and } x = (x_1, x_2) \in \mathcal{E}$$

Consider the discounted problem with discount factor $0 \leq \alpha < 1$:

$$v^\gamma(x) = \max_\sigma \max_\tau \mathbb{E}\left[\sum_{k=0}^{\tau-1} \alpha^k r(U_k, X_k) + \alpha^\tau \gamma \mid X_0 = x\right] \quad ,$$

where the maximization holds over all strategies σ and all stopping times τ (with respect to the filtration of the history process associated to σ). Write the dynamic programming equation satisfied by v^γ.

One can alternatively consider a Markov decision process with enlarged state space $\bar{\mathcal{E}} = \mathcal{E} \cup \{0\}$ (0 is a cemetery point), enlarged action space $\bar{\mathcal{C}} = \mathcal{C} \cup \{0\}$, constrained action spaces given by $\bar{\mathcal{C}}(x) = \bar{\mathcal{C}}$ when $x \in \mathcal{E}$ and $\bar{\mathcal{C}}(0) = \{0\}$, transition probabilities extending M by $\overline{M}^{(u)}_{x,x'} = M^{(u)}_{x,x'}$ for $x, x' \in \mathcal{E}$ and $u \in \mathcal{C}$, and $\overline{M}^{(0)}_{x,0} = 1$ for $x \in \bar{\mathcal{E}}$, and instantaneous reward $\bar{r}$ extending r by $\bar{r}(u, x) = r(u, x)$ for $x \in \mathcal{E}$ and $u \in \mathcal{C}$, $\bar{r}(0, x) = \gamma$ for $x \in \mathcal{E}$ and $\bar{r}(u, 0) = 0$ for $u \in \bar{\mathcal{C}}$. Then v^γ is the restriction to $\mathcal{E}$ of the value of the discounted problem with infinite horizon and discount factor α, and the action $u = 0$ means stopping.

Q 2.3. Build an optimal policy (or pure Markov stationary strategy) π of the problem by using the dynamic programming equation of Q.2.2.

Q 2.4. For each $i \in \{1, 2\}$, let $v^{i,\gamma}$ be the value function of the stopping time problem :

$$v^{i,\gamma}(x_i) = \max_\tau \mathbb{E}\left[\sum_{k=0}^{\tau-1} \alpha^k r_i(Y^i_k) + \alpha^\tau \gamma \mid Y^i_0 = x_i\right] \quad ,$$

where the maximization holds over all stopping times τ and $x_i \in \mathcal{E}_i$. Let $F^{i,\gamma}$ be the operator from $\mathbb{R}^{\mathcal{E}_i}$ to itself such that

$$[F^{i,\gamma}(v)](x_i) = \max\left(\gamma\ ,\ r_i(x_i) + \alpha \sum_{x'_i \in \mathcal{E}_i} M^i_{x_i x'_i} v(x'_i)\right) \quad .$$

Write the dynamic programming equation satisfied by $v^{i,\gamma}$ using $F^{i,\gamma}$.

Q 2.5. Using the properties of dynamic programming operators, deduce that, for all $x \in \mathcal{E}_i$, the map $\gamma \in \mathbb{R} \mapsto v^{i,\gamma}(x) \in \mathbb{R}$ is convex. (Note that this implies that it is continuous.)

Q 2.6. Denote by $\| \cdot \|_\infty$ the sup-norm on $\mathbb{R}^{\mathcal{E}_i}$. Show that

$$\|v^{i,\gamma}\|_\infty \leq \max(|\gamma|, \frac{\|r_i\|_\infty}{1-\alpha}) \quad .$$

Q 2.7. Let

$$m_i(x_i) = \min\{\gamma \mid r_i(x_i) + \alpha \sum_{x'_i \in \mathcal{E}_i} M^i_{x_i x'_i} v^{i,\gamma}(x'_i) \leq \gamma\} \quad .$$

Describe the optimal policy for the problem of Q.2.4 when $m_i(x_i) > \gamma$. Using the above properties of $\gamma \mapsto v^{i,\gamma}(x_i)$, describe also the optimal policy when $m_i(x_i) < \gamma$.

We want to infer the optimal policy of the problem of Q.2.2 using the optimal policies of each problem of Q.2.4 with $i = 1, 2$.

Q 2.8. Show that for all $x \in \mathcal{E}$, and $i = 1, 2$, $v^{i,\gamma}(x_i) \leq v^\gamma(x_1, x_2)$.

Q 2.9. Deduce that if $x \in \mathcal{E}$ is such that $\gamma < m_i(x_i)$ for some $i = 1, 2$, then it is not optimal to stop in state x, that is for any optimal policy $\pi : \mathcal{E} \to \bar{\mathcal{C}}$, we have $\pi(x) \neq 0$.

Q 2.10. Let F^γ be the operator from $\mathbb{R}^\mathcal{E}$ to itself such that

$$[F^\gamma(v)](x) = \max\left([F^{1,\gamma}(v(\cdot, x_2))](x_1)\ ,\ [F^{2,\gamma}(v(x_1, \cdot))](x_2)\right) \quad ,$$

where $v(x_1, \cdot)$ denotes the map w from $\mathcal{E}_2 \to \mathbb{R}$ such that $w(x_2) = v(x_1, x_2)$ for all $x_2 \in \mathcal{E}_2$. Show that the function

$$w(x) = v^{1,\gamma}(x_1) + v^{2,\gamma}(x_2) - \gamma$$

satisfies $F^\gamma(w) \leq w$. Deduce that $v^\gamma \leq w$.

Q 2.11. Deduce that if $x \in \mathcal{E}$ is such that $\max(m^1(x_1), m^2(x_2)) \leq \gamma$, then it is optimal to stop ($\pi(x) = 0$).

Q 2.12. Deduce also that if $x \in \mathcal{E}$ is such that $m^2(x_2) \leq \gamma < m^1(x_1)$, then it is optimal to choose the action 1 ($\pi(x) = 1$). One may need to use that $v^{i,\gamma}(x_i) \leq v^\gamma(x_1, x_2)$ for all $x \in \mathcal{E}$.

---

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

We consider a stationnary Markov Decision Process with finite state space $\mathcal{E} = \{1, \cdots, n\}$ and control space $\mathcal{C}$. We assume that $\mathcal{C}(x) = \mathcal{C}$ is independent of the state, and denote by $M_{xy}^{(u)}$ the transition probabilities (formally, $M_{xy}^{(u)} = \mathbb{P}\left(X_{n+1} = y \mid X_n = x, U_n = u\right)$, for $x, y \in \mathcal{E}$ and $u \in \mathcal{C}$). $\mathbb{R}_+$ denotes the set of positive reals. We consider a positive function $\gamma : \mathcal{E} \times \mathcal{C} \to \mathbb{R}_+$ which can be seen either as a discount factor, a multiplicative cost or a multiplicative reward.

Given a (pure or random) strategy $\sigma = (\sigma_k)_{k \geq 0}$, we consider

$$J^{(T,\sigma)}(x) := \mathbb{E}\left[\prod_{k=0}^{T-1} \gamma(X_k, U_k) \mid X_0 = x\right] \quad , \tag{1}$$

$$\zeta^\sigma(x) := \limsup_{T \to \infty} \left\{J^{(T,\sigma)}(x)\right\}^{\frac{1}{T}} \quad , \tag{2}$$

where the expectation and the process $(X, U) := (X_k, U_k)_{k \geq 0}$ are induced by σ. The following study is related to the problem of maximization or minimization of the ergodic risk sensitive criterion $\zeta^\sigma(x)$ among all strategies.

We denote by Π the set of all stationary (feedback) policies, that is the maps $\pi : \mathcal{E} \to \mathcal{C}$. For any $\pi \in \Pi$, we denote by $M^{(\pi)}$ the $\mathcal{E} \times \mathcal{E}$ matrix with entry (x, y) equal to $M_{xy}^{(\pi(x))}$ and by $A^{(\pi)}$ the $\mathcal{E} \times \mathcal{E}$ matrix with entry (x, y) equal to $\gamma(x, \pi(x)) M_{xy}^{(\pi(x))}$. (Recall that the elements of $\mathbb{R}^\mathcal{E}$ are seen either as functions from $\mathcal{E}$ to $\mathbb{R}$ or as (column) vectors, in particular as elements of $\mathbb{R}^n$.)

In the sequel, we denote by Exp the map from $\mathbb{R}^\mathcal{E}$ to $\mathbb{R}_+^\mathcal{E}$ which takes the exponential componentwise : $\mathrm{Exp}(v) = (\exp(v_x))_{x \in \mathcal{E}}$, for $v = (v_x)_{x \in \mathcal{E}} \in \mathbb{R}^\mathcal{E}$. We also denote by Log the inverse map of Exp, so $\mathrm{Log}(v) = (\log(v_x))_{x \in \mathcal{E}}$, for $v = (v_x)_{x \in \mathcal{E}} \in \mathbb{R}_+^\mathcal{E}$.

Q 3.1. For any $\pi \in \Pi$, write the Kolmogorov equation satisfied by the functions $J^{(T,\pi)}$, with $T \geq 0$, and deduce that $\zeta^\pi(x) \leq \rho(A^{(\pi)})$, for all $x \in \mathcal{E}$, where ρ denotes the spectral radius of a matrix.

Q 3.2. We assume in this question that the graph of $M^{(\pi)}$ is strongly connected (or equivalently that $M^{(\pi)}$ is irreducible). Using the existence of a Perron vector of $A^{(\pi)}$ (a positive eigenvector associated to the eigenvalue $\rho(A^{(\pi)})$), show that there exists $C > 0$ such that $J^{(T,\pi)}(x) \geq C\rho(A^{(\pi)})^T$, for all $x \in \mathcal{E}$. Deduce that $\zeta^\pi(x) = \rho(A^{(\pi)})$, for all $x \in \mathcal{E}$.

3.1 The minimization problem

Q 3.3. Consider the operator $\mathcal{B}$ from $\mathbb{R}_+^{\mathcal{E}}$ to itself defined as follows :

$$[\mathcal{B}(v)]_x := \min_{u \in \mathcal{U}} \left\{ \gamma(x,u) \sum_{y \in \mathcal{E}} M^u_{x,y} v_y \right\} \quad ,$$

and let $\mathcal{B}^T$ be the Tth iterate of this operator. When the map $v \in \mathbb{R}_+^{\mathcal{E}}$ is fixed, show that $[\mathcal{B}^T(v)]_x$ is the value of a Markov Decision problem with the above MDP parameters and a finite horizon criterion to be precised.

Q 3.4. Let $\mathcal{T} = \mathrm{Log} \circ \mathcal{B} \circ \mathrm{Exp} : \mathbb{R}^{\mathcal{E}} \to \mathbb{R}^S$, $v \mapsto \mathrm{Log}(\mathcal{B}(\mathrm{Exp}(v)))$. Show that $\mathcal{T}$ is order preserving and additively homogeneous.

Q 3.5. Deduce that, for all $\alpha < 1$, the operator $\mathcal{T}_\alpha$ such that $\mathcal{T}_\alpha(v) = \mathcal{T}(\alpha v)$ is a contraction on $\mathbb{R}^{\mathcal{E}}$ and has a unique fixed point, that shall be denoted by v_α.

Q 3.6. Let $L := \max_{x,u} |\log \gamma(x,u)|$. Show that $(1-\alpha)\|v_\alpha\|_\infty \leq L$, where for any $v \in \mathbb{R}^{\mathcal{E}}$, $\|v\|_\infty = \max_{x \in \mathcal{E}} |v_x|$ denotes the sup-norm.

Q 3.7. Since $\mathcal{E}$ is a finite set, there exists z_α such that $v_\alpha(z_\alpha) = \min_{x \in \mathcal{E}} v_\alpha(x)$. Then, we set

$$\mu_\alpha = (1-\alpha) v_\alpha(z_\alpha) \quad \text{and} \quad w_\alpha(x) = v_\alpha(x) - v_\alpha(z_\alpha).$$

Show that

$$\exp(\mu_\alpha + w_\alpha(x)) = \min_{u \in \mathcal{U}} \left\{ \gamma(x,u) \sum_{y \in \mathcal{E}} M^u_{x,y} \exp\left(\alpha w_\alpha(y)\right) \right\} \quad .$$

For all $\alpha < 1$, we shall consider $\pi_\alpha \in \Pi$ such that $\pi_\alpha(x)$ realizes the minimum in the previous equation ($\pi_\alpha(x)$ exists since $\mathcal{C}$ is a finite set).

Q 3.8. Using the properties that $\mathcal{E}$ and $\mathcal{C}$ are finite sets, and the previous results, show that for any sequence $(\alpha_n)_{n \in \mathbb{N}}$ in $[0,1)$ converging to 1, there exists a subsequence also denoted $(\alpha_n)_{n \in \mathbb{N}}$ satisfying :

$$\pi_{\alpha_n}(x) = \pi(x) \ , \quad z_{\alpha_n} = z \ , \quad \forall n \in \mathbb{N} \ , \quad \lim_{n \mapsto \infty} \mu_{\alpha_n} = \mu \ , \quad \text{and} \ \lim_{n \mapsto \infty} w_{\alpha_n}(x) = w(x) \ , \quad \forall x \in \mathcal{E} \ .$$

for some $\pi \in \Pi$, $z \in \mathcal{E}$, $\mu \in \mathbb{R}$ and some map $w : \mathcal{E} \to [0, +\infty]$ which may be infinite.

Q 3.9. Show that μ and w satisfy the following equations :

$$\exp(\mu + w_x) = \min_{u\in\mathcal{U}} \left\{ \gamma(x,u) \sum_{y\in\mathcal{E}} M^u_{x,y} \exp(w_y) \right\} = \sum_{y\in\mathcal{E}} A^{(\pi)}_{x,y} \exp(w_y) \ \ \forall x \in \mathcal{E} \ \ .$$

Q 3.10. Let $\pi \in \Pi$, $z \in \mathcal{E}$ and $w \in [0, +\infty]^{\mathcal{E}}$ be as in Q. 3.8 and let $\mathcal{I} = \{x \mid w(x) < +\infty\}$. Show that $z \in \mathcal{I}$ and that $\mathcal{I}$ satisfies the following invariance property :

$(I(\pi))$ If $x \in \mathcal{I}$ and $M^{(\pi)}_{x,y} > 0$ then $y \in \mathcal{I}$.

Q 3.11. Assume now that for all $\pi \in \Pi$, the matrix $M^{(\pi)}$ is irreducible. Show that any set $\mathcal{I} \subset \mathcal{E}$ satisfying the invariance property $(I(\pi))$ of Q. 3.10 for some $\pi \in \Pi$ is either empty or equal to $\mathcal{E}$. Deduce that the map w of Q. 3.8 is finite everywhere, $w \in \mathbb{R}^{\mathcal{E}}$, and that it satisfies $\mu\mathbf{1} + w = \mathcal{T}(w)$.

Q 3.12. Let $v^* = \mathrm{Exp}(w)$, $\lambda^* = \exp(\mu)$ and π be as in Q. 3.8. Show that $A^{(\pi)}v^* = \lambda^* v^*$ and $A^{(\pi')}v^* \geq \lambda^* v^*$, for all $\pi' \in \Pi$, that $\mathcal{B}(v^*) = \lambda^* v^*$ and that π is an optimal policy in the computation of $\mathcal{B}(v^*)$ in Q. 3.3.

Q 3.13. Deduce, from the previous question and using Perron-Frobenius theorem, the following equalities :

$$\lambda^* = \min_{\pi\in\Pi} \rho(A^{(\pi)}) = \sup\{\lambda \mid \lambda > 0,\ v \in \mathbb{R}^{\mathcal{E}}_+,\ \text{s.t. } A^{(\pi')}v \geq \lambda v \ \forall \pi' \in \Pi\} \ \ .$$

3.2 The maximization problem

All the above arguments can be done similarly when the minimization is replaced by maximization in the definition of $\mathcal{B}$, leading, under the irreducibility of all matrices $M^{(\pi)}$, to the existence of $\lambda^* \in \mathbb{R}$ and $v^* \in \mathbb{R}^{\mathcal{E}}_+$ such that $\mathcal{B}(v^*) = \lambda^* v^*$ and

$$\lambda^* = \max_{\pi\in\Pi} \rho(A^{(\pi)}) = \inf\{\lambda \mid \lambda > 0,\ v \in \mathbb{R}^{\mathcal{E}}_+,\ \text{s.t. } A^{(\pi')}v \leq \lambda v \ \forall \pi' \in \Pi\} \ \ .$$

Q 3.14. We can obtain without any assumption $\mu \in \mathbb{R}$, $w \in [0, +\infty]^{\mathcal{E}}$ and $\pi \in \Pi$ such that

$$\exp(\mu + w_x) = \max_{u\in\mathcal{U}} \left\{ \gamma(x,u) \sum_{y\in\mathcal{E}} M^u_{x,y} \exp(w_y) \right\} = \sum_{y\in\mathcal{E}} A^{(\pi)}_{x,y} \exp(w_y) \ \ \forall x \in \mathcal{E} \ \ .$$

Show that, for all $\pi' \in \Pi$, the set $\mathcal{I} = \{x \mid w(x) < +\infty\}$ satisfies the invariance property $(I(\pi'))$. Deduce that a sufficient condition for w to be finite is now that the graph of the MDP is strongly connected.

Q 3.15. We admit the following result (see Lemma 5.62 of Lecture notes) : for every $v \in \mathbb{R}^{\mathcal{E}}$ and probability ν on $\mathcal{E}$, we have

$$\log\left(\sum_{x'\in\mathcal{E}} \nu_{x'} \exp(v_{x'}) \right) = \sup_{\theta\in\Delta_S} \left(-\mathcal{KL}(\theta,\nu) + \sum_{x'\in\mathcal{E}} \theta_{x'} v_{x'} \right) \ \ ,$$

where $\Delta_{\mathcal{E}}$ is the set of probabilities on $\mathcal{E}$ and $\mathcal{KL}$ is the *Kullback-Leibler distance* defined as :

$$\mathcal{KL}(\theta,\theta') = \sum_{x\in\mathcal{E}} \theta_x \log\left(\frac{\theta_x}{\theta'_x}\right) \ \ .$$

Show that λ^* and v^* can be computed by solving a Linear Program with an infinite number of linear inequality constraints, or equivalently by solving a convex program.