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 12 novembre 2024
Durée 3h

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.

Recall that this exam is based on all Lectures 1 to 9 (Tuesday Sept. 10 to Tuesday Nov. 5, 2024), and is for Master 2 students or ENSTA students who need to obtain more ECTS.

1 Problem 1 : an optimal stopping time problem

Let $N \in \mathbb{N} \setminus \{0\}$ be fixed and $(X_k)_{k\in[1,N]}$ a finite sequence of positive independent random variables with the same law q, taking their values in a finite subset $\mathcal{E}$ of $(0, +\infty)$. Consider the following game which is played in at most N steps. We assume that the player knows the law q. At the nth step of the game, he observes the value of X_n and then decides to stop the game forever or to continue to play. If he stops at step n, then he will obtain the value of X_n as his payoff (there are no other payoffs). The aim of the player is to maximize the expected value of this payoff. Let $v_n(x)$ be the optimal value for the game, when at step n, the player knows that $X_n = x$.

Q 1.1. Write the recursive equation which is satisfied by the sequence of functions $v_n(x)_{n\in[1,N]}$.

Q 1.2. Let $G_n = \mathbb{E}\left[v_n(X_n)\right]$, show that $G_n = \phi(G_{n+1})$ for some function ϕ, and give the expression of $\phi(x)$.

Q 1.3. Show that the function ϕ is non-negative and non-decreasing and that $\phi(x) - x$ is non-negative, non-increasing (w.r.t. x), and tends to zero when x goes to infinity.

Q 1.4. Using the sequence $(G_n)_{n=1,N}$, give the expression of the optimal stopping rule and comment the strategy.

2 Problem 2 : a stopping time problem with partial observation

Consider the same game as in Problem 1, but now, at the ith step of the game, the player only observes the value of $Y_n = f(X_n)$ for some map $f : \mathcal{E} \to \mathcal{Y}$, and then using all its pas observations, he decides if he stops the game forever or if he continues to play. The aim of the player is still to maximize the expected value of his payoff.

Q 2.1. Recall that an optimal stopping time problem can be replaced by a MDP with control space $\mathcal{C} = \{0, 1\}$, where $u = 1$ means that the player is stopping and $u = 0$ that he continues, and enlarged the state space $\mathcal{E}' = \mathcal{E} \cup \{c\}$ where c is a cemetery point and apropriate reward.

Show that the above problem can be written as a finite horizon problem for a partially observable Markov Decision Process (POMDP) with state space $\mathcal{E}'$, control space $\mathcal{C}$ and observation space $\mathcal{Y}' = \mathcal{Y} \cup \{c\}$ (and with $f(c) = c$). Give the transition probabilities.

Q 2.2. Write the dynamic programming equation satisfied by the value $v_n(b)$ of the POMDP as a function of the belief b at step n.

Q 2.3. For each $y \in \mathcal{Y}$, let us denote by $b^{(y)}$ the probability vector on $\mathcal{E}'$ with support included in $\mathcal{E}$, and such that $b_x^{(y)} = \mathbb{P}(X_0 = x \mid f(X_1) = y)$. For $y = c$, let us denote by $b^{(c)}$ the Dirac measure at point c of $\mathcal{E}'$ (that is the probability measure on $\mathcal{E}'$ with support $\{c\}$).

Show that the only reachable beliefs (after one step) of the POMDP are the vectors $b^{(y)}$ with $y \in \mathcal{Y}'$, and that the above equation restricted to these vectors can be interpreted to a Dynamic Programming equation of an optimal stopping time problem on the state space $\mathcal{Y}$.

Q 2.4. Write the Dynamic Programming equation satisfied by $G_n = \sum_{y \in \mathcal{Y}} \mathbb{P}(f(X_0) = y)\, v_n(b^{(y)})$. Comment.

3 Problem 3 : Value iterations for ergodic problems

We consider a stationary Markov Decision Process $(X_k)_{k \geq 0}$ over a finite state $\mathcal{E}$, with finite actions spaces $\mathcal{C}(x) \subset \mathcal{C}$ in each state $x \in \mathcal{E}$ and transition probability vectors $M_x^{(u)}$ over $\mathcal{E}$, $M_x^{(u)} = \left(M_{xy}^{(u)}\right)_{y \in \mathcal{E}}$, for all state $x \in \mathcal{E}$ and action $u \in \mathcal{C}(x)$, and instantaneous reward functions $r : \mathcal{E} \times \mathcal{C} \to \mathbb{R}$. The aim of the problem is to maximize the following mean-payoff (average time, ergodic) criteria

$$J_x((X_k)_{k \geq 0}; (U_k)_{k \geq 0}) = \limsup_{T \to \infty} \left\{ \frac{1}{T} \mathbb{E}\left[\sum_{k=0}^{T} r(X_k, U_k) \mid X_0 = x \right] \right\} \ ,$$

over all the stategies $\sigma = (\sigma_k)_{k \geq 0}$ (with U_k following this strategy). We are looking for optimal stategies σ that are stationnary (feedback) policies, that is $\sigma = (\sigma_k)_{k \geq 0}$, with $\sigma_k = \pi$ for all $k \geq 0$ and $\pi \in \Pi := \{\pi : \mathcal{E} \to \mathcal{C} \mid \pi(x) \in \mathcal{C}(x),\ \forall x \in \mathcal{E}\}$ (and so $U_k = \pi(X_k)$).

We denote by $\mathcal{B} : \mathbb{R}^{\mathcal{E}} \to \mathbb{R}^{\mathcal{E}}$,

$$[\mathcal{B}(v)]_x = \sup_{u \in \mathcal{C}(x)} \left(r(x, u) + M_x^{(u)} v \right) \ ,$$

the dynamic programming operator of the problem.

In all the problem, we assume that there exists $\rho^* \in \mathbb{R}$ and $v^* \in \mathbb{R}^{\mathcal{E}}$ such that

$$\rho^* \mathbf{1} + v^* = \mathcal{B}(v^*) \ , \tag{1}$$

where $\mathbf{1}$ is the unit vector of $\mathbb{R}^{\mathcal{E}}$, $\mathbf{1} = (1)_{x \in \mathcal{E}}$. We also fix a parameter $\gamma \in (0, 1)$, and construct the operator

$$\mathcal{B}_\gamma :\ v \in \mathbb{R}^{\mathcal{E}} \mapsto \mathcal{B}_\gamma(v) = (1 - \gamma) v + \mathcal{B}(\gamma v) \ .$$

Q 3.1. Using the operator $\mathcal{B}_\gamma$, and (1), write an equation satisfied by ρ^* and $w^* = v^*/\gamma$, and explain to which MDP with mean-payoff this equation is related.

Q 3.2. Explain how (1) or the equation of Q 3.1 can help in solving the initial mean-payoff problem ?

We denote by $\|\cdot\|_\infty$ the sup-norm on $\mathbb{R}^{\mathcal{E}}$, and consider the following nonsymmetric and not nonnegative "seminorm" on $\mathbb{R}^{\mathcal{E}}$:

$$\mathbf{t}\,[v] = \max\{v(x) \mid x \in \mathcal{E}\} \ ,$$

so that we have

$$\|v\|_\infty = \max\{|v(x)| \mid x \in \mathcal{E}\} = \max(\mathbf{t}\,[v]\,, \mathbf{t}\,[-v]) \ .$$

Q 3.3. Show that $\mathcal{B}$ is nonexpansive for the seminorm $\mathbf{t}\,[\cdot]$, that is

$$\mathbf{t}\left[\mathcal{B}(v) - \mathcal{B}(v')\right] \leq \mathbf{t}\left[v - v'\right] \quad \forall v, v' \in \mathbb{R}^{\mathcal{E}} \ .$$

and that the same holds for $\mathcal{B}_\gamma$.

We shall now consider the value iteration algorithm associated to $\mathcal{B}_\gamma$ starting from the zero vector :

$$w_0 = 0, \quad w_{n+1} = \mathcal{B}_\gamma(w_n) \in \mathbb{R}^{\mathcal{E}}, \quad n \geq 0 \ .$$

Q 3.4. Show that $\mathbf{t}\,[w_{n+1} - w_n]$ and $\mathbf{t}\,[w_n - w_{n+1}]$ are nonincreasing sequences (with respect to n), and that $-\mathbf{t}\,[w_n - w_{n+1}] \leq \mathbf{t}\,[w_{n+1} - w_n]$. Deduce that both sequences are converging.

Let us denote by ρ^+ the limit of $\mathbf{t}\,[w_{n+1} - w_n]$, and by ρ^- the limit of $-\mathbf{t}\,[w_n - w_{n+1}]$.
For any $z \in \mathcal{E}$, we denote

$$\rho_n^{(z)} = w_n(z) - w_{n-1}(z) \in \mathbb{R}$$
$$h_n^{(z)} = w_n - w_n(z)\mathbf{1} \in \mathbb{R}^{\mathcal{E}}, \quad \text{that is } h_n^{(z)}(x) = w_n(x) - w_n(z) \quad \forall x \in \mathcal{E} \ .$$

Q 3.5. Show that there exists some state z^+ and an increasing sequence $(n_k)_{k\geq 0}$ of integers such that $\mathbf{t}\,[w_n - w_{n-1}] = \rho_n^{(z^+)}$ holds for all integers $n = n_k$, $k \geq 0$. In particular $\rho_{n_k}^{(z^+)}$ converges towards ρ^+ when k goes to infinity.

Q 3.6. Show that

$$\rho_{n+1}^{(z)} \leq (1-\gamma)\rho_n^{(z)} + \gamma\mathbf{t}\,[w_n - w_{n-1}]$$

holds for all $n \geq 1$. Deduce that $\rho_{n_k-1}^{(z^+)}$ also converges towards ρ^+ when k goes to infinity.

Q 3.7. Show that the same holds for $\rho_{n_k-p}^{(z^+)}$ for any positive integer p.

Q 3.8. Write an equation satisfied by $\rho_n^{(z)}$, $h_n^{(z)}$ and $h_{n-1}^{(z)}$, for all n.

Q 3.9. Recall that the sequence w_n satisfies for all n :

$$\|w_n - n\rho^* \mathbf{1}\|_\infty \leq 2\|w^*\|_\infty \ .$$

Show that the sequence $h_n^{(z)}$ is bounded and deduce that there exists a constant C such that, for all $n \geq p \geq 1$, we have

$$|\rho_n^{(z^+)} + \cdots + \rho_{n-p+1}^{(z^+)} - p\,\rho^*| \leq C \ .$$

Q 3.10. Deduce that $\rho^+ = \rho^*$ and so $\lim_{n\to\infty} \mathbf{t}\,[w_{n+1} - w_n] = \rho^*$.

Q 3.11. We shall admit the symmetrical property $\lim_{n\to\infty} \mathbf{t}\,[w_n - w_{n+1}] = -\rho^*$. Deduce that for all $z \in \mathcal{E}$,

$$\lim_{n\to\infty} \rho_n^{(z)} = \rho^*$$
$$\lim_{n\to\infty} h_{n+1}^{(z)}(x) - h_n^{(z)}(x) = 0 \quad \forall x \in \mathcal{E} \ .$$

Q 3.12. Show that any limit point h of the sequence $(h_n^{(z)})_{n\geq 0}$ satisfies

$$\rho^* + h = \mathcal{B}_\gamma(h) \ , \quad h(z) = 0 \ . \tag{2}$$

For any policy π, we denote by $r^{(\pi)}$, $M^{(\pi)}$ and $\mathcal{B}^{(\pi)}$ the reward vector, Markov transition matrix and operator of the Markov decision problem with fixed policy π :

$$r_x^{(\pi)} = r(x, \pi(x)) \ , \quad M^{(\pi)} = (M_{xy}^{(\pi(x))})_{x,y\in\mathcal{E}} \ , \quad \mathcal{B}^{(\pi)}(v) = r^{(\pi)} + M^{(\pi)} v \ .$$

Assume that all the matrices $M^{(\pi)}$, with $\pi \in \Pi$, have a unique final (recurrence) class and that z belongs to this final class.

Q 3.13. Given two solutions $h, h' \in \mathbb{R}^{\mathcal{E}}$ of (2), show that there exist a policy $\pi \in \Pi$ such that

$$h - h' \leq M^{(\pi)}(h - h') \ .$$

Q 3.14. Deduce that $h \leq h'$, that the solution h to (2) is unique and conclude to the convergence of the above value iteration.

Q 3.15. Is there another algorithm computing the solution of (2), under the same assumptions ?