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

Marianne Akian

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

Mardi 14 novembre 2023
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 (Tuesday Sept. 12 to Wednesday Nov. 8, 2023), and that its purpose is to validate the master course. For master students who are also ENSTA students, it can be used to validate ENSTA Course SOD312 (possibly with more ECTS).

1 Problem 1

A manufacturer is producing a product which uses a raw material with high volatility (random variation of price), so, in order to avoid big variations of its profit, he is managing this raw material by bying some amount of it in advance at a lower price or for selling it later at a higher price. We shall denote by $\phi(p)$ the every-day profit of production of the product by the firm when the price of the raw material is equal to p. We assume that the price at time (day) $n \in \mathbb{N}$ of the raw material is equal to P_n and that $(P_n)_{n\in\mathbb{N}}$ is a sequence of independent random variables with values in some finite set $\mathcal{P}$ and with the same law which is known by the firm. Each time (day) $n \in \mathbb{N}$, the manufacturer can choose to bye a certain amount $U_n \in \mathbb{Z}$ of the raw material, independent of its current needs, where $U_n < 0$ means that he is indeed selling the amount $-U_n$. This induces a cost P_nU_n. The amount U_n need however to satisfy the constraint that the total amount X_n of raw material which is not needed currently, and is thus put in stock, is between 0 and $\bar{x}$.

Q 1.1. The manufacturer wants to maximize the total expected return during N days. Write this problem as a Markov decision process with state space $\mathcal{P} \times \mathcal{X}$, with $\mathcal{X} = \{0, \dots, \bar{x}\}$, action space $\mathbb{Z}$, and finite horizon N, and precise the dynamics and criterion.

Q 1.2. Write the dynamic programming equation allowing one to compute the value function of this problem and explain how optimal strategies of the manufacturer can be obtained.

Q 1.3. The manufacturer wants now to maximize its expected amount of profit by time unit in the long run. Write this problem as a MDP with mean-payoff (long run time-average payoff) criterion.

Q 1.4. Relate the value of the above problem with the existence of a solution (ρ, v) to the equation :

$$\rho + v(p,x) = \max_{u\in\mathbb{Z},\ x+u\in\mathcal{X}} \left(\phi(p) - pu + \mathbb{E}\left[v(P_1, x+u)\right]\right), \quad \forall (p,x) \in \mathcal{P}\times\mathcal{X}$$

where ρ is a real scalar.

Q 1.5. Show the existence of a solution to the previous equation.

Q 1.6. Let ρ and v be as above and denote $w(x) = \mathbb{E}\left[v(P_1, x)\right]$ for $x \in \mathcal{X}$. Deduce an ergodic equation for w.

Q 1.7. Solve the equation of w, find ρ as a function of ϕ and of the law of P_1, and find a stationary optimal policy.

2 Problem 2

Consider a stationary Markov Decision Process with finite state space $\mathcal{E}$, finite control spaces $\mathcal{C}$ be finite sets, and $\mathcal{C}(x) \subset \mathcal{C}$, for $x \in \mathcal{E}$, and transition probability vectors $(M^{(u)}_{xy})_{y\in\mathcal{E}}$, for all $(x,u) \in \mathcal{E}\times\mathcal{C}$ (so that $M^{(u)}_{xy} \geq 0$ and $\sum_{y\in\mathcal{E}} M^{(u)}_{xy} = 1$). Let $r : \mathcal{E}\times\mathcal{C} \to \mathbb{R}$ be a reward map, and consider the map $\mathcal{B}$ from $\mathbb{R}^{\mathcal{E}}$ to itself defined by :

$$[\mathcal{B}(v)](x) = \sup_{u\in\mathcal{C}(x)} \left(r(x,u) + \sum_{y\in\mathcal{E}} M^{(u)}_{xy} v(y) \right), \tag{1}$$

We 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 above Markov decision process, and for each $\pi \in \Pi$, we denote by $r^{(\pi)} \in \mathbb{R}^{\mathcal{E}}$ the vector with entries $r^{(\pi)}_x = r(x, \pi(x))$, by $M^{(\pi)} \in \mathbb{R}^{\mathcal{E}\times\mathcal{E}}$ the Markov matrix with entries $M^{(\pi)}_{xy} = M^{(\pi(x))}_{xy}$, and by $\mathcal{B}^{(\pi)}$ the affine operator :

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

We shall assume that all the Markov matrices $M^{(\pi)}$ are ergodic meaning that they have a unique final class. We shall also assume that there exists $c \in \mathcal{E}$ such that c belongs to the final class of each $M^{(\pi)}$.

Given $\pi_0 \in \Pi$, we construct the sequences $\pi^k \in \Pi$ and $v^k \in \mathbb{R}^{\mathcal{E}}$, for $k \geq 0$, respectively of policies and value functions of the following variant of the policy iteration algorithm for the ergodic case :

1. $\rho^{(k)}$ and v^k satisty $\rho^{(k)} + v^k = \mathcal{B}^{(\pi^k)}(v^k)$ and $v^k(c) = 0$.
2. π_{k+1} is an optimal policy for v^k, meaning that $\mathcal{B}(v^k) = \mathcal{B}^{(\pi_{k+1})}(v^k)$ that is

$$\pi_{k+1}(x) \in \operatorname*{Argmax}_{u\in\mathcal{C}(x)} \left(r(x,u) + \sum_{y\in\mathcal{E}} M^{(u)}_{xy} v^k(y) \right) \quad \text{for all } x \in \mathcal{E} \ ,$$

such that

$$\pi_{k+1}(x) = \pi_k(x) \quad \text{if } \pi_k(x) \in \operatorname*{Argmax}_{u\in\mathcal{C}(x)} \left(r(x,u) + \sum_{y\in\mathcal{E}} M^{(u)}_{xy} v^k(y) \right) \quad \text{for all } x \in \mathcal{E} \ .$$

Q 2.1. Show that there exists a solution $(\rho, v) \in \mathbb{R} \times \mathbb{R}^{\mathcal{E}}$ to the following equation

$$\rho + v(x) = [\mathcal{B}(v)](x) \quad x \in \mathcal{E} \ , \tag{2}$$

and explain why the solution ρ is unique. We shall denote by ρ^* the unique solution.

In the sequel, $\|\cdot\|$ denote the sup-norm on $\mathbb{R}^{\mathcal{E}}$: $\|v\| = \max_{x\in\mathcal{E}} |v(x)|$. For any Markov matrix M, we denote by $M_{(c)}$ the matrix obtained from M by putting to zero all entries of M that are in column c (so $M_{(c)}$ is not a Markov matrix). We assume that there exists a vector $\varphi \in \mathbb{R}^{\mathcal{E}}$ with positive coordinates such that

$$\varphi \geq \mathbf{1} + M_{(c)}^{(\pi)}\varphi \ , \forall \pi \in \Pi \ , \tag{3}$$

and let K be a bound on its coefficients, $K \geq \|\varphi\|$.

Q 2.2. Let L be the map which associates to any couple $(\rho, v) \in \mathbb{R} \times \mathbb{R}^{\mathcal{E}}$ such that $v_c = 0$, the vector $w \in \mathbb{R}^{\mathcal{E}}$ such that $w(x) = v(x)\varphi(x)^{-1} + \rho$. Show that L is a (linear) bijection.

Q 2.3. Show that (2) is equivalent to the fixed point equation for w of an operator $\widetilde{\mathcal{B}}$ on $\mathbb{R}^{\mathcal{E}}$, $\widetilde{\mathcal{B}}(w) = w$, and show that $\widetilde{\mathcal{B}}$ is the Bellman operator of a new MDP on the same state and action spaces, but with a state dependent discount factor equal to $\alpha(x) = 1 - 1/\varphi(x)$, and new parameters that will be precised.

Q 2.4. Show that $\widetilde{\mathcal{B}}$ is a contracting operator.

Q 2.5. Show that the construction of Q 2.3 can also be applied to the operators $\mathcal{B}^{(\pi)}$. Explain the relation between the above policy iteration and the policy iteration for the fixed point of $\widetilde{\mathcal{B}}$.

Q 2.6. Can we deduce a contraction of the error on $\rho^{(k)}$ and v^k ?

3 Problem 3

Consider a provider (of electricity or telecommunication networks) with $N-1$ different offers (contracts), and assume that N corresponds to an alternate offer from concurent providers, which is fixed. Denote $\mathcal{E} = \{1, \ldots, N\}$. The provider has also a finite set $\mathcal{C}$ of actions on his offers, for instance the set of different prices during the day, or the week.

There are many customers, all of the same type and undistinguishable, so the provider only consider at each time or period t the proportion $b_n(t)$ of customers using the offer n, so that $b(t) = (b_n(t))_{n\in\mathcal{E}} \in \Delta_{\mathcal{E}}$, the set of probability vectors over $\mathcal{E}$. His return at each period or time t is then linear with respect to these proportions :

$$r(u, b) = \sum_{m=1}^{N} R_m(u) b_m \ ,$$

wher $R_m(u) > 0$ for $m < N$ and $R_N(u) = 0$, and the provider wishes to maximize his total payoff on the period of time $[0, T]$:

$$\sum_{t=0}^{T-1} r(u(t), b(t)) \ .$$

All customers have the same probability to switch from offer n to offer m which depends on the action $u \in \mathcal{C}$ of the provider and has the form :

$$M_{nm}^{(u)} = \beta P_m^{(u)} + (1-\beta)\delta_{nm}$$

where $\delta_{nm} = 1$ if $n = m$ and 0 otherwise, $P^{(u)}$ is a probability vector on $\mathcal{E}$, and $\beta \in (0, 1]$ is the probability to switch. Here $P_m^{(u)}$ can be seen as a measure of the utility of any customer to use offer m, and β is the probability of switching.

Seeing the proportions of customers in each offer as probabilities, we approximate the dynamics of the proportions by using the Fokker-Plank equation that is :

$$b_n(t+1) = \sum_{m \in \mathcal{E}} b_m(t) M_{mn}^{(u(t))} \ .$$

Q 3.1. Write this problem as a finite horizon problem for a deterministic control problem over the state space $\Delta_{\mathcal{E}}$. One can also consider the finite subsets S_t obtained by considering only all the possible values of $b(t)$ when $b(0)$ is fixed and the controls $u(s)$, with $s \in \{0, \dots, T-1\}$, are taken in the action space $\mathcal{C}$.

Q 3.2. Write the dynamic programming equation satisfied by the value function $b \in \Delta_{\mathcal{E}} \mapsto v_k(b)$, for $k = 0, \dots, T$, and explain why it is related to a partially observable Markov Decision Process. Precise the parameters of the latter.

Q 3.3. Assume now that $\beta = 1$. Show that when $t \geq 1$, $b(t)$ takes only a finite number of possible values.

Q 3.4. Under the same assumption that $\beta = 1$, we consider the discounted infinite horizon problem with discount factor $0 \leq \alpha < 1$, which consists in maximizing

$$\sum_{t=0}^{\infty} \alpha^t r(u(t), b(t)) \ ,$$

over all possible strategies. Write the corresponding dynamic programming equation. Deduce that there is a periodic sequence of controls $(u_k)_{k \geq 0}$ with period $K \leq N$, such that for every initial proportion $b(0)$, there exists an optimal (open-loop) control sequence which coincides with the periodic sequence $(u_k)_{k \geq 0}$ after some finite time $t \leq N$.