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

Marianne Akian

ENSTA Course SOD312

Mercredi 25 octobre 2023
Durée 3h

Problems 1 and 2 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 Lectures 1 to 5 (Tuesday Sept. 12 to Tuesday Oct. 10, 2023), and that its only purpose is to validate ENSTA Course SOD312. (Another exam will be proposed later to Master students.)

1 Problem 1

The maintenance of roads of a region is assigned to a company under a contract for a fixed period of T years. The state of the roads during the year $t \in \{1, \dots, T\}$ is measured by an integer $X_t \in [[N]] := \{0, \dots, N\}$. A larger X_t means that the roads are in worse condition.

The contract includes a fixed price C paid by the region to the company for the entire period (of T years) with a penalty paid by the company to the region each year $t \in \{1, \dots, T\}$ in which the roads are above some fixed level $l < N$. This penalty is (to simplify) equal to the gap $\max(X_t - l, 0)$ between the state and the level.

Each year t, the company decides to repair or not the roads. We denote by $U_t \in \{0, 1\}$ this choice, $U_t = 1$ meaning repairing and $U_t = 0$ not repairing.

The cost of repairing the roads is $\gamma + \delta X_t$, that is it includes a fixed cost $\gamma > 0$ and a proportional one, with proportional factor $\delta > 0$. Repaired roads have necessarily a measure of state equal to 0.

Q 1.1. The first aim of the company is to maximize the total return of the contract, that is

$$R((X_t)_{t\geq 0}, (U_t)_{t\geq 0}) = C - \sum_{k=1}^{T} [\max(X_t - l, 0) + U_t(\gamma + \delta X_t)] \ . \tag{1}$$

Assume first that when roads are not repaired the state of the roads is increasing regularly each year : $X_{t+1} = X_t + 1$, except when in the worst state. Explain why the above maximization problem is a deterministic optimal control problem (deterministic Markov decision problem) with additive criteria, finite horizon T, and precise the dynamics.

Q 1.2. Denote by $v^T(x)$ the maximum of the criterion (1) as a function of the initial state of roads ($X_1 = x$). Write the dynamic programming equation allowing one to compute the function v^T and find optimal strategies of the company in the form of feedback policies.

Q 1.3. The contract can be renewed only under the condition that the worst level of the roads during the period, that is $\max(X_1, \ldots, X_T)$, is below some fixed level $L < N$. In order to take into account the possible loss of the contract, the company decides to maximize the sum of the total return of the contract and of some increasing function ϕ of the gap to the level :

$$J((X_t)_{t\geq 0}, (U_t)_{t\geq 0}) = C - \left(\sum_{k=1}^{T} [\max(X_t - l, 0) + U_t(\gamma + \delta X_t)]\right) + \phi(L - \max(X_1, \ldots, X_T)) \ .$$

Write this maximization problem as a deterministic optimal control problem with enlarged state space.

Q 1.4. Give a recurrence equation allowing one to find the new value function w^T and optimal strategies.

Q 1.5. We come back to criterion (1) of Q 1.1, but we assume now that, when roads are not repaired the state of the roads is increasing randomly as in $X_{t+1} = \min(X_t + W_t, N)$ where $(W_t)_{t\geq 0}$ is a sequence of i.i.d. random variables with values in $[[N]]$. We shall denote by $p(w)$ the probability of the event $W_t = w$. Write the dynamic programming equation for the value function v^T associated to the maximization of $\mathbb{E}\left[R((X_t)_{t\geq 0}, (U_t)_{t\geq 0})\right]$ and give the optimal strategies.

Q 1.6. Show that, for all $T \geq 0$, v^T is a nonincreasing function of $x \in [[N]]$.

Q 1.7. Consider the following operators from $\mathbb{R}^{[[N]]}$ to itself :

$$[\mathcal{B}^{(0)}(v)](x) := -\max(x - l, 0) + \sum_{w\in[[N]]} p(w)\, v(\min(x + w, N))$$
$$[\mathcal{B}^{(1)}(v)](x) := -\max(x - l, 0) - (\gamma + \delta x) + v(0) \ .$$

Show that if the map $\mathcal{B}^{(0)}(v^T) - \mathcal{B}^{(1)}(v^T)$ is nonincreasing for some $T = T_0$ then it is nonincreasing for all $T \geq T_0$.

Q 1.8. Assume that the condition of Q 1.7 is satisfied for some T_0 and let

$$L_t^* := \max\left\{x \mid [\mathcal{B}^{(0)}(v^t)](x) \geq [\mathcal{B}^{(1)}(v^t)](x)\right\} \ .$$

Show that, when $k \leq T - T_0$, the optimal policy at time k is to repair the roads if $X_k \geq L_{T-k}^*$.

---

2 Problem 2 : on the complexity of policy iterations

Let $\mathcal{E}$ and $\mathcal{C}$ be finite sets, $0 < \alpha < 1$ be a scalar, $\mathcal{C}(x)$ be subsets of $\mathcal{C}$ associated to $x \in \mathcal{E}$, $g : \mathcal{E} \times \mathcal{C} \to \mathbb{R}$ be a map, and for all $(x, u) \in \mathcal{E} \times \mathcal{C}$, let $(M_{xy}^{(u)})_{y\in\mathcal{E}}$ be the entries of some row probability vector (so that $M_{xy}^{(u)} \geq 0$ and $\sum_{y\in\mathcal{E}} M_{xy}^{(u)} = 1$). Consider the map $\mathcal{B}$ from $\mathbb{R}^{\mathcal{E}}$ to itself defined by :

$$[\mathcal{B}(v)](x) = \sup_{u\in\mathcal{C}(x)} \left(g(x, u) + \alpha \sum_{y\in\mathcal{E}} M_{xy}^{(u)} v(y)\right) \ .$$

Q 2.1. To which Markov decision problem corresponds a fixed point $v \in \mathbb{R}^{\mathcal{E}}$ of $\mathcal{B}$? What is in that case the meaning of the above parameters α, g, $M^{(u)}_{xy}$ and of v ? How can we find an optimal policy for this problem ?

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 $g^{(\pi)} \in \mathbb{R}^{\mathcal{E}}$ the vector with entries $g^{(\pi)}_x = g(x, \pi(x))$, 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 :

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

Let $\pi_0 \in \Pi$ and 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 policy iteration algorithm, that we next recall :

1. v^k is a fixed point of $\mathcal{B}^{(\pi^k)}$,
2. π_{k+1} is an optimal policy for v^k, meaning that $\mathcal{B}(v^k) = \mathcal{B}^{(\pi_{k+1})}(v^k)$.

Let v^* denote the (unique) fixed point of $\mathcal{B}$ and $\|\cdot\|$ denote the sup-norm on $\mathbb{R}^{\mathcal{E}}$: $\|v\| = \max_{x\in\mathcal{E}} |v(x)|$. The following questions use the monotonicity properties of the operators $\mathcal{B}$ and $\mathcal{B}^{(\pi)}$, and of the sequence $(v^k)_{k\geq 0}$ shown in the lectures.

Q 2.2. Show that

$$\|v^{k+1} - v^*\| \leq \alpha \|v^k - v^*\|, \quad \text{for all } k \geq 0 \ .$$

Q 2.3. Consider the cost $c(x, u) = v^*(x) - g(x, u) - \alpha \sum_{y\in\mathcal{E}} (M^{(u)}_{xy} v^*(y))$, and for all $\pi \in \Pi$, denote by $c^{(\pi)}$ the vector with entries $c^{(\pi)}_x = c(x, \pi(x))$. Show that $c^{(\pi)} = v^* - \mathcal{B}^{(\pi)}(v^*)$, and give an interpretation of $\|c^{(\pi)}\|$ as a measure of the distance of π to an optimal policy.

Q 2.4. Show that $0 \leq c^{(\pi^k)} \leq v^* - v^k$, for all $k \geq 0$.

Q 2.5. Show that, for all $k \geq 0$, $\|v^k - \mathcal{B}^{(\pi^k)}(v^*)\| \leq \alpha \|v^k - v^*\|$ and deduce that $\|v^k - v^*\| \leq \frac{1}{1-\alpha} \|c^{(\pi^k)}\|$.

Q 2.6. Deduce that there exist an integer p and a positive constant $\mu < 1$ such that, for all $t \geq k+p$,

$$\|c^{(\pi^t)}\| \leq \mu \|c^{(\pi^k)}\| \ .$$

Denote by $\mathcal{A} = \{(x, u) \in \mathcal{E} \times \mathcal{C} \mid u \in \mathcal{C}(x)\}$ and denote by $\mathcal{G}(\pi) \subset \mathcal{A}$ the graph of any map $\pi \in \Pi$.

Q 2.7. Assume that π^k is not optimal, and let (x, u) realize the maximum of $c(x, u)$ on the graph of π^k. Show that, for all $t \geq k + p$, with p as in Q 2.6, (x, u) cannot belong to the graph of π^t.

Q 2.8. Deduce that there is a sequence of subsets $\mathcal{A}^k$ of $\mathcal{A}$ such that $\mathcal{G}(\pi^k) \subset \mathcal{A}^k$ for all $k \geq 0$, and such that $\mathcal{A}^{kp}$ is (strictly) decreasing as far as $\pi^{(k-1)p}$ is not optimal. Deduce that the number of policy iterations is bounded from above by :

$$k_{\max} := (m - n)(1 + \lfloor \frac{\log(1-\alpha)}{\log(\alpha)} \rfloor) \ ,$$

where m is the cardinality of $\mathcal{A}$.