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 9 novembre 2021
Durée 3h

This text contains 2 different exams :

M2 Exam consists in Problems 1 and 4 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 or 3 for these students.

ENSTA Exam consists in Problems 1, 2 and 3 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 4 for these students (moreover this problem may use notions that were not teached to ENSTA students).

Problem 2 is a following of Problem 1, but Problems 1, 3 and 4 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)

We consider a Paris taxicab driver. For the present problem 1, he is driving passengers inside Paris only. The taxi driver can get a trip either by being hailed, or by waiting at a cab station, or by radio. For each demand of a trip by a potential passenger, the benefit is known from the begining (it may be proportional to its length in kilometers), a minimal duration is also known, but the actual duration maybe increased by traffic jam. Given the destination of the trip, the benefit and the minimal duration, the taxi driver can accept to do it or not.

Let us "discretize" a working day of the taxi driver into N steps (corresponding to time units), and assume that at each step $n \in \{0, \dots, N-1\}$ (or time interval $[n, n+1)$) in which the taxi is free, a passenger is asking for a trip to the taxi driver. We shall denote by B_n and D_n the benefit and minimal duration of this trip.

We assume that all the variables D_n, B_n are defined for all $n \geq 1$, and that they constitute (all together) a sequence of identically distributed independent random variables (note however that D_n and B_n should be dependent). We also assume that (D_n, B_n) take a finite set of positive integer and positive real values respectively, and let (d_k, b_k), $k = 1, \dots K$ be their possible values, and p_k be their probabilities. We also denote by $\bar{d}$ the maximum of all the d_k. As said above, the minimal duration of a trip is known from the begining, however the variations in duration arrive during the

trip. We shall model these variations as follows : once the trip is accepted by the driver, at each time step, the duration may increase of one time unit with some probability $\gamma \in (0,1)$ (which may depend on the current duration), or stay unchanged with probability $1-\gamma$. Moreover, the taxi driver can accept a trip until the last step, although he will have to drive after step N.

The aim of the taxi driver is to maximize his expected total income (in one day).

Q 1.1. For each step $n \geq 1$, if the taxi is free and the driver choose to accept a demand of a passenger, his benefit will be equal to B_n and the minimal time he will be busy is the minimal duration of the trip, that is D_n. Otherwise, he has to look for a new passenger and this costs him some fixed amount C.

For each step $n \geq 1$, let $Y_n \in \{0, \ldots, \bar{d}\}$ be the (minimal) remaining time units before the taxi gets free, with $Y_n = 0$ if the taxi is already free. We assume that there exists a nondecreasing function $\gamma : \mathbb{R}_+ \to [0,1)$ such that if $Y_n > 0$, then the remaining time at step $n+1$ will be $Y_n - 1$ with probability $1-\gamma(Y_n)$ (the total duration of the trip is unchanged) or stays Y_n with probability $\gamma(Y_n)$ (the total duration of the trip increases of one time unit). Denote by U_n the choice of the taxi driver : $U_n = 1$ if the driver accepts the demand of the passenger and $U_n = 0$ if he does not accept it, and in particular if the taxi is already occuped.

Formulate the taxi driver problem as the maximization of an expected additive reward over a finite horizon N, for the Markov decision process with state process (Y_n, B_n, D_n) and control process U_n. Describe the dynamics of the controlled process and the instantaneous rewards.

Q 1.2. Deduce that the value function of the problem satisfies the following Dynamic programming equation, for $y \in \{0, \ldots, \bar{d}\}$ and $(b,d) \in \{(b_k, d_k) \mid k = 1, \ldots, K\}$,

$$v_n(y,b,d) = \sum_{k=1}^{K} p_k(\gamma(y)v_{n+1}(y,b_k,d_k) + (1-\gamma(y))v_{n+1}(y-1,b_k,d_k)) \ , \quad \text{if } y > 0$$

$$v_n(0,b,d) = \max(-C + \sum_{k=1}^{K} p_k v_{n+1}(0,b_k,d_k), b + \sum_{k=1}^{K} p_k(\gamma(y)v_{n+1}(d,b_k,d_k) + (1-\gamma(y))v_{n+1}(d-1,b_k,d_k)))$$

$$v_T(y,b,d) = 0 \ .$$

Q 1.3. Denote $w_n(y) = \sum_{k=1}^{K} p_k v_n(y,b_k,d_k)$, for $y \in \{0, \ldots, \bar{d}\}$. Write a recurrence equation for the functions w_n.

Q 1.4. Show that the optimal policy is given by a threshold, that is if the taxi is free at step n ($Y_n = 0$), then the driver accepts the passenger if $B_n \geq \bar{b}(n, D_n)$. Give the expression of the threshold.

Q 1.5. Consider the case $\gamma \equiv 0$. Write a recurrence equation for $z_m = w_{N-m}(0)$ and show that the map $m \mapsto z_m$ is convex on any interval $(-\infty, n]$ (by induction on $n \geq 0$). Deduce that the above threshold $\bar{b}(n,b)$ is nonincreasing with respect to n. Explain.

---

2 Problem 2 (to validate the ENSTA lectures only)

We consider a variant of Problem 1.

Now, the Paris taxicab driver is driving passengers either inside Paris or between Orly Airport and Paris or Roissy Airport and Paris, in both sides. The price for a trip between Airports and Paris is fixed and one shall assume that the benefit of the driver for such a trip is fixed. This trip has a minimal duration, which maybe increased by traffic jam.

When in Paris, the taxi driver can get a trip inside Paris, but also to one of the Airports, he can refuse it. In Airports, the taxi driver need to go to the cab station, and cannot refuse a demand of a trip between this airport and Paris.

Let us denote by O, R, P the possible positions of the taxi cab, where O is for Orly, R for Roissy, and P for Paris, and denote by $X_n \in \{O, R, P\}$ the position of the taxi cab, at step n, or the position at the end of its current trip, when it is busy. For each $n = 1, \dots, N$, if the taxi cab is free and in Airport $A \in \{O, R\}$ ($X_n = A$), then the passenger is asking for a trip between this airport and Paris, we denote by B^A the fixed benefit of the trip for the taxi driver, and by D^A the minimal duration of the trip (in time units). If the taxi cab is free and in Paris ($X_n = P$), we denote by $Z_n \in \{O, R, P\}$ the destination of the (potential) passenger who is asking for a trip. If the destination Z_n is one of the airports, then again B^A will be the fixed benefit of the trip for the taxi driver, and D^A will be the minimal duration of the trip. Otherwise, if $Z_n = P$, B_n and D_n denote as before the benefit and minimal duration of the trip.

We assume now that $(Z_n, D_n, B_n)_{n\geq 1}$ is a sequence of identically distributed independent random variables taking its values in $\{(z_k, d_k, b_k) \mid k = 1, \dots K\}$, and that p_k is the probability of (z_k, d_k, b_k). When $z_k = A \in \{O, R\}$, the variables d_k and b_k are fixed to D^A and B^A. We shall thus assume that $z_1 = O$, $z_2 = R$ and $z_k = P$ for $k \geq 3$. We also model the evolution of the duration of the trips (in Paris or between Paris and the airports) as in Problem 1.

The aim of the taxi driver is still to maximize his expected total income (in one day).

Q 2.1. At each step n in which the taxi cab is free in Paris, the driver has the choice of either accept the trip which is proposed and gain B_n, refuse it and stay in Paris at the cost C, or go to one of the airports $A \in \{O, R\}$ at the cost C^A. The corresponding values of U_n can be denoted $1, 0, O, R$.

Formulate the new problem as the maximization of an expected additive reward over a finite horizon N, for the Markov decision process with state process $(X_n, Y_n, Z_n, B_n, D_n)$ and control process U_n. Describe the new dynamics of the controlled process and the instantaneous rewards.

Q 2.2. Write the corresponding dynamic programming equation of the value function $v_n(x, y, z, b, d)$, and deduce the one of $w_n(x, y)$, where $w_n(x, y) = \sum_{k=1}^{K} p_k v_n(x, y, z_k, b_k, d_k)$ when $x = P$ and $w_n(A, y) = v_n(A, y, P, B^A, D^A)$ for $A \in \{O, R\}$.

Q 2.3. Instead of bounding the duration N of the day, we assume now that the taxi driver is becoming more and more tired, and thus at each step n his interest in the following benefits of the day is multiplied by some factor $\alpha \in (0, 1)$. Another way to see this is that conditionnally to take a passenger at time n, and to finish the trip at time $n + m$, the taxi driver will be returning to home after the trip with a probability equal to $1 - \alpha^m$. Formulate the new problem either as the maximization of an expected discounted payoff over an infinite horizon or by adding a cemetery state c to the state space of the MDP.

Q 2.4. Write the corresponding dynamic programming equation of the value function $v(x, y, z, b, d)$. Denote $w(x, y) = \sum_{k=1}^{K} p_k v(x, y, z_k, b_k, d_k)$ when $x = P$ and $w(x, y) = v(A, y, P, B^A, D^A)$ when $x = A \in \{O, R\}$. Show that w satisfies, for $x \in \{P, O, R\}$ and $y \in \{0, \dots, \bar{d}\}$,

$$w(x, y) = \alpha(\gamma(y)w(x, y) + (1 - \gamma(y))w(x, y - 1)) \ , \quad \text{if } y > 0$$

$$w(P, 0) = \sum_{k=1}^{K} p_k \left[\max(-C + \alpha w(P, 0), -C^O + w(O, D^O), -C^R + w(R, D^R), b_k + w(z_k, d_k))\right]$$

$$w(A, 0) = b^A + w(P, D^A) \ , \quad \text{for } A \in \{O, R\} \ .$$

Q 2.5. Explain why this equation has a unique solution. What is the policy iteration algorithm computing w and v ? How many steps are needed for such an algorithm in general, given the number of possible actions ?

Q 2.6. Show that a stationary optimal policy is given by a threshold, and give the expression of the threshold.

Q 2.7. Assume that $\gamma \equiv 0$. Give a necessary and sufficient condition for the optimal policy to be to accept all demands ?

---

3 Problem 3 (to validate the ENSTA lectures only)

Q 3.1. Let $(W_n)_{n\geq 0}$ be a sequence of independent identically distributed random variables taking its values in a subset $\mathcal{W}$ of $\mathbb{R}$ with zero expectation and finite variance.

Consider the Markov Decision Process with (infinite) state space $\mathcal{E} = \mathbb{R}$, action (control) space $\mathcal{C} = \mathbb{R}$, and dynamics

$$X_{n+1} = \lambda X_n + U_n + W_n$$

with $\lambda \in (0, 1)$. We want to solve the finite horizon Markov Decision problem :

$$v^T(x) = \min \mathbb{E}\left[\left(\sum_{k=0}^{T-1} \alpha^k c(X_k, U_k)\right) + \alpha^T \varphi(X_T) \mid X_0 = x\right] \ ,$$

in which $\alpha \in (0, 1)$, $c(x, u) = (ax^2 + u^2)/2$ and $\varphi(x) = x^2/2$, for some $a > 0$. Write a Dynamic programming equation associated to this problem. (note that assuming that all parameters are rational, one can reduce the problem to the countable state space $\mathcal{E} = \mathbb{Q}$ and so all the results of the course can be applied).

Q 3.2. Write a recurrence equation in T for v^T.

Q 3.3. Show that, for all $T \geq 0$, the value function satisfies : $v^T(x) = a_T x^2/2 + b_T$, for some $a_T > 0$ and $b_T \in \mathbb{R}$.

Q 3.4. Show that an optimal feedback policy at time k (when the horizon is T) can be of the form $\pi_k(x) = -c_k x$ with $c_k > 0$.

Q 3.5. What happens when T goes to infinity ?

---

4 Problem 4 (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 $\mathcal{C}$. We denote by $\mathcal{A} := \{(x,u) \mid x \in \mathcal{E},\ u \in \mathcal{C}(x)\}$ the set of all possible pairs (state, action), and 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 process.

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

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 :

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

Q 4.1. Explain for which Markov Decision Processes $\mathcal{B}$ is the dynamic programming operator (explain the parameters). Interpret also $\mathcal{B}^{(\pi)}$ as a dynamic programming operator, and explain why

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

Q 4.2. For all $v \in \mathbb{R}^{\mathcal{E}}$, we denote $\mathsf{t}(v) := \max_{x \in \mathcal{E}} v(x)$, $\mathsf{b}(v) := \min_{x \in \mathcal{E}} v(x)$, and $\|v\|_H = \mathsf{t}(v) - \mathsf{b}(v)$, the latter beeing called Hilbert's semi-norm. Show that, for all $v, w \in \mathbb{R}^{\mathcal{E}}$, we have

$$\mathsf{t}(\mathcal{B}(v) - \mathcal{B}(w)) \leq \max_{(x,u) \in \mathcal{A}} \sum_{y \in \mathcal{E}} M_{xy}^{(u)} (v - w)(y)\ .$$

Deduce that $\mathcal{B}$ is nonexpansive for the Hilbert semi-norm $\| \cdot \|_H$:

$$\|\mathcal{B}(v) - \mathcal{B}(w)\|_H \leq \|v - w\|_H \quad \forall v, w \in \mathbb{R}^{\mathcal{E}}\ .$$

(on can first show that $\mathcal{B}$ is non expansive for t and b).

Consider the following constant :

$$\beta := 1 - \min_{(x,u),(x',u') \in \mathcal{A}} \left[\sum_{y \in \mathcal{E}} \min(M_{xy}^{(u)}, M_{x'y}^{(u')}) \right] \in [0,1]\ . \tag{1}$$

Q 4.3. Deduce from Question 4.2 the following inequality :

$$\|\mathcal{B}(v) - \mathcal{B}(w)\|_H \leq \beta \|v - w\|_H \quad \forall v, w \in \mathbb{R}^{\mathcal{E}}\ .$$

(One may have to use the set of the $y \in \mathcal{E}$ in which $M_{xy}^{(u)} < M_{x'y}^{(u')}$.)

Q 4.4. For $\alpha \in (0,1)$, let $\mathcal{B}_\alpha : v \mapsto \mathcal{B}(\alpha v)$. Deduce from the previous question that $\mathcal{B}_\alpha$ is contracting for the Hilbert's semi-norm with contraction factor $\alpha\beta$.

Q 4.5. Denote by v^* the fixed point of $\mathcal{B}_\alpha$, and let v^n be the sequence of value iterations for $\mathcal{B}_\alpha$, starting in $v^0 \in \mathbb{R}^{\mathcal{E}}$. Show that

$$\|v^n - v^*\|_H \leq \frac{(\alpha\beta)^n}{1-\alpha\beta}\|v^1 - v^0\|_H \ .$$

Q 4.6. For all $v \in \mathbb{R}^{\mathcal{E}}$, we denote by $\|v\|_\infty = \max(\mathsf{t}(v), \mathsf{b}(v))$ the sup-norm of v. Show that for all $v, w \in \mathbb{R}^{\mathcal{E}}$, and $\pi \in \Pi$, we have

$$\mathsf{t}(\mathcal{B}^{(\pi)}(v) - \mathcal{B}^{(\pi)}(w)) \leq \max_{(x,u)\in\mathcal{A}} \sum_{y\in\mathcal{E}} M^{(u)}_{xy}(v-w)(y) \ .$$

Deduce that, for all $v, w \in \mathbb{R}^{\mathcal{E}}$, we have

$$\|\mathcal{B}(v) - \mathcal{B}^{(\pi)}(v) - \mathcal{B}(w) + \mathcal{B}^{(\pi)}(w)\|_\infty \leq \beta\|v-w\|_H \ .$$

Denote by Π^* the set of all optimal stationnary policies obtained from Dynamic programming equation, or equivalently the set of elements π of Π such that $\mathcal{B}(\alpha v^*) = \mathcal{B}^{(\pi)}(\alpha v^*)$. Assume that Π^* is a proper subset of Π. This implies that there exists a positive real $\kappa > 0$ such that for all $\pi \in \Pi \setminus \Pi^*$, and $x \in \mathcal{E}$, we have

$$\text{either} \quad v^*(x) - \mathcal{B}^{(\pi)}(\alpha v^*)(x) \geq \kappa \quad \text{or} \quad v^*(x) - \mathcal{B}^{(\pi)}(\alpha v^*)(x) = 0 \ .$$

Q 4.7. For the sequence v^n of value iterations for $\mathcal{B}_\alpha$, starting in $v^0 \in \mathbb{R}^{\mathcal{E}}$, we shall consider

$$N_0 := \inf\{k \geq 0 \mid \alpha\beta\|v^n - v^*\|_H < \kappa \quad \forall n \geq k\} \ .$$

Show that for all $n \geq N_0$, any optimal policy π for v^n ($\mathcal{B}(\alpha v^n) = \mathcal{B}^{(\pi)}(\alpha v^n)$) belongs to Π^*.

Q 4.8. Does it means that the sequence v^n converges in finite time ? If not, how can we compute v^* using v^n ?

Q 4.9. Give an upper bound on N_0 using v^0, $\alpha\beta$ and κ only.

Q 4.10. If $\beta < 1$, what does the above result says for the relative value iterations for the ergodic equation ?