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

Marianne Akian and Jean-Philippe Chancelier

ENSTA SOD312 / M2 Mathématiques et Applications, U. Paris Saclay, Optimization

Mardi 12 novembre 2019
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 ofthe other problems 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).

The problems are independent (except for the statements of Problems 1 and 2) and often questions can be solved without solving the previous questions. The solution can be written either in French or English.

1 Problem 1 (for all students)

An unscrupulous innkeeper charges a different rate for a room as the day progresses, depending on wether he has many or few vacancies. His objective is to maximize his expected total income during the day. Let x be the number of empty rooms at the start of the day, and let y be the number of customers that will ask for a room during the day.

We assume that y is known to the inkeeper. When a customer arrives the innkeeper proposes a price $q \in \{q_1, \dots, q_n\}$ where $0 < q_1 < \dots < q_n$. The customer will accept the offer q_i with probability p_i and refuse the offer with probability $(1 - p_i)$. If the customer refuses the offer, he won't come back during the day.

Q 1.1. Let t corresponds to the t-th arrival. Denote by $X_t \in \{0, \dots, x\}$ the number of empty rooms between the $(t-1)$-th and the t-th arrivals of a customer and by $Y_t \in \{0, \dots, y\}$ the remaining number of customers which will arrive after the same period of time. Formulate the innkeeper problem as the maximization of an expected additive reward over an infinite horizon, for the controlled Markov decision process (X_t, Y_t), in which the instananeous reward evaluates to zero when X_t or Y_t is equal to zero. Describe the control process and the dynamics of the controlled process.

Q 1.2. Deduce that the value function v satisfies the following Dynamic programming equation

$$v(x,y) = \max_{i \in \{1,\dots,n\}} \Big(p_i\big(q_i + v(x-1, y-1)\big) + (1-p_i) v(x, y-1) \Big)$$

Q 1.3. Denoting by $T = y$ the number of customers which will arrive during an entire day, reformulate the previous optimization problem as a finite horizon Markov decision process with state process $(X_t)_{t\geq 0}$, and time horizon T, and write the corresponding Dynamic Programming equation.

Q 1.4. Assume that $p_1 q_1 < \ldots < p_n q_n$ and that $p_1 > p_2 > \ldots > p_n$. Show that the innkeeper should always charge at the highest rate q_n.

---

2 Problem 2 (to validate the ENSTA lectures only)

We consider a variant of Problem 1 in which the customer proposes a price q_i with probability p_i. The innkeeper can accept or refuse the offer. If he refuses the offer, then the customer won't come back during the day.

Q 2.1. Consider an additional state variable Q_t of the controlled Markov chain which is the price proposed by the t-th customer. Show in that case that the Dynamic Programming equation satisfied by the value function v maximizing the innkeeper profits is

$$v(x, y, q) = \max\big(q + \sum_{i=1}^{n} p_i v(x-1, y-1, q_i), \sum_{i=1}^{n} p_i v(x, y-1, q_i)\big).$$

Q 2.2. Denote $w(x, y) = \sum_{i=1}^{n} p_i v(x, y, q_i)$. Write a fixed point equation for w.

Q 2.3. Show that the optimal policy is given by a threshold, that is the offer of the customer is accepted if it is larger than a function $\overline{q}(x, y)$. Give the expression of the threshold.

Q 2.4. Show by induction on y, that $x \in \mathbb{N} \mapsto w(x, y)$ is nondecreasing.

Q 2.5. Show by induction on y, that $x \in \mathbb{N} \mapsto w(x, y)$ is convex, that is the slope $w(x, y) - w(x-1, y)$ is a nondecreasing function of $x \in \mathbb{N}$.

Q 2.6. Deduce that $x \mapsto \overline{q}(x, y)$ is nondecreasing and give an interpretation of this property.

Q 2.7. Assume now that y is not known and random, and that after each arrival of a customer, the probability of an additional arrival is $\alpha \in (0, 1)$. This means that with probability $1-\alpha$ there will be no arrivals anymore.

Consider now a state process composed of X_t, the number of empty rooms, Q_t, the price proposed by the t-th customer, and Y_t, the possibility of an additional arrival, that is $Y_t = 0$ if there will be no arrivals in the future and $Y_t = 1$ otherwise. Formulate the new innkeeper problem as the maximization of an expected additive reward over an infinite horizon, for the controlled Markov decision process (X_t, Q_t, Y_t), in which the instananeous reward evaluates to zero when X_t or Y_t is equal to zero.

Q 2.8. Show that the value function satisfies the following Dynamic Programming equation

$$v(x, q, 1) = \max\big(q + \alpha \sum_{i=1}^{n} p_i v(x-1, q_i, 1), \alpha \sum_{i=1}^{n} p_i v(x, q_i, 1)\big).$$

with $v(x, q, 0) = 0$ for all x, q.

Q 2.9. Interpret this equation as the Dynamic Programming equation of an infinite horizon discounted problem, and explain why this equation has a unique solution.

Q 2.10. How can it be solved ?

Q 2.11. Denote $w(x) = \sum_{i=1}^{n} p_i v(x, q_i, 1)$. Using the properties of the fixed point equation in 2.8, show that w satisfies the same properties as above, that is that $x \in \mathbb{N} \mapsto w(x, y)$ is nondecreasing and convex.

Q 2.12. Deduce that the optimal policy is still given by a threshold, and that the threshold function $x \mapsto \overline{q}(x)$ is nondecreasing.

---

3 Problem 3 (to validate the ENSTA lectures only)

We consider a discrete time controlled Markov chain (or Markov decision process) whose dynamics is given by

$$X_{t+1} = f_t(X_t, U_t, W_{t+1}) \ .$$

We denote by $\mathbb{X}$ the state space of the controlled Markov chain, by $\mathbb{U}$ its action (or control) space and assume that the disturbances W_{t+1} belong to the set $\mathbb{W}$. In addition to the sequence $(X_t)_{t\geq 0}$, the decision maker observes along time a sequence $\big(Y_t\big)_{t\geq 0}$ of random variables taking values in $\mathbb{Y} = \{1, \ldots, n\}$, $Y_t : \Omega \to \{1, \ldots, n\}$ is such that $\mathbb{P}(\{Y_t = k\}) = \pi_k$ for $k \in \mathbb{Y}$. The sequence of random variables $(W_{t+1}, Y_t)_{t\geq 0}$ are independent (and independent of X_0). The law of W_{t+1} knowing Y_t is given and known and denoted by $\mathbb{P}_{t+1}(w|y)$ (we assume that it is a discrete law over $\mathbb{W}$). We want to compute

$$\min_{U \in \mathcal{U}} \mathbb{E}\Big[\sum_{i=0}^{T-1} c_i(X_i, U_i, W_{i+1}) + K_T(X_T) \Big] \ , \tag{1}$$

where the strategies/policies depend possibly on all the history, that is on the past states $(X_s, Y_s)_{s\leq t}$ and the past controls $(U_s)_{s<t}$.

Q 3.1. Explain which state you have to use, why you can use state feedbacks and propose a Dynamic Programming Equation to compute the solution to Problem (1), in which the value functions are functions of $(x, y) \in \mathbb{X} \times \mathbb{Y}$.

Q 3.2. Show that the mapping $\overline{V}_t(x) := \sum_{i=1}^{n} \pi_i V_t(x, i)$ satisfies a recursive equation similar to Dynamic Programming.

Q 3.3. Rewritting the previous equation, by using a control depending on the value of the observation $y \in \mathbb{Y}$, show that this equation can be interpreted as the Dynamic Programming equation of a controlled Markov chain with state space $\mathbb{X}$, control space $\mathbb{U}^n$ and disturbance space $\mathbb{Y} \times \mathbb{W}$.

---

4 Problem 4 (to validate the full M2 lectures only)

Let us consider a MDP over the state space $\mathcal{E} = \{1, \dots, n\}$, with action space $\mathcal{C}(x) \subset \mathcal{C} = \mathcal{E}$ when the current state is $x \in \mathcal{E}$ and deterministic dynamics $X_{k+1} = U_k$, meaning that the transition probabilities are independent of time and equal to $P(X_{k+1} = y \mid X_k = x, U_k = u) = M_{xy}^{(u)} = \delta_{yu}$ (where $\delta_{xy} = 1$ if $x = y$ and 0 otherwise). Let $r : \mathcal{E} \times \mathcal{C} \to \mathbb{R}$ be a reward function.

We consider the maximization of the following long run time average criteria, among all state and control processes $(X_k, U_k)_{k\geq 0}$ determined by any strategy and starting at some state $x \in \mathcal{E}$:

$$J((X_k)_{k\geq 0}; (U_k)_{k\geq 0}) = \limsup_{T\to\infty} \left\{ \frac{1}{T} \mathbb{E}\left[\sum_{k=0}^{T} r(X_k, U_k) \right] \right\} \quad , \tag{2}$$

and denote by $\zeta(x)$ its value (its supremum).

We associate to the above process the directed graph $\mathcal{G}$, with set of nodes equal to $\mathcal{E}$ and set of arcs $\mathcal{A}$ equal to the set of $(x, y) \in \mathcal{E} \times \mathcal{E}$ such that $y \in \mathcal{C}(x)$.

Q 4.1. We assume that $\mathcal{G}$ is strongly connected.

Using results of the course, show that there exists $\rho \in \mathbb{R}$ and $v \in \mathbb{R}^{\mathcal{E}}$ such that

$$\rho + v(x) = F_x(v) := \max_{y\in\mathcal{C}(x)} (r(x, y) + v(y)) \quad \forall x \in \mathcal{E} \ .$$

Q 4.2. Deduce the value of $\zeta(x)$, for all $x \in \mathcal{E}$.

Q 4.3. Let π be a deterministic policy (or feedback strategy), that is an element of $\Pi := \{\pi : \mathcal{E} \to \mathcal{C} \mid \pi(x) \in \mathcal{C}(x), \ \forall x \in \mathcal{E}\}$, and consider (following the course) the vector and matrix

$$r_x^{(\pi)} = r(x, \pi(x)), \quad M_{xy}^{(\pi)} = M_{xy}^{\pi(x)} = \delta_{y\pi(x)}, \quad \forall x, y \in \mathcal{E} \ .$$

Show that the graph of the Markov matrix $M^{(\pi)}$ necessarily contains one cycle, that is a path $(x_1, \dots, x_k, x_{k+1})$ for some $k \leq n$, such that $x_{k+1} = x_1$ and $x_i \neq x_j$ when $1 \leq i \neq j \leq k$.

Q 4.4. Show that the set $C = \{x_1, \dots, x_k\}$ of nodes of this cycle is a final class of the Markov matrix $M^{(\pi)}$.

Q 4.5. Let m_C be the probability measure over $\mathcal{E}$ which is equal to the uniform probability over C. Show that m_C is an invariant measure of $M^{(\pi)}$.

Q 4.6. Recall that $F(v) \geq r^{(\pi)} + M^{(\pi)} v$ for all $v \in \mathbb{R}^{\mathcal{E}}$ and $\pi \in \Pi$. Deduce that

$$\rho \geq \frac{r(x_1, x_2) + \cdots + r(x_{k-1}, x_k) + r(x_k, x_1)}{k} \ .$$

Q 4.7. Show that indeed the previous inequality holds for all cycles $(x_1, \dots, x_k, x_1)$ of $\mathcal{G}$.

Q 4.8. Let π be an optimal policy for a solution v of the ergodic equation in Q 4.1, that is a policy such that

$$\rho \mathbf{1} + v = r^{(\pi)} + M^{(\pi)} v \ .$$

Show that

$$\rho = \frac{r(x_1, x_2) + \cdots + r(x_{k-1}, x_k) + r(x_k, x_1)}{k} \ ,$$

for any cycle $(x_1, \dots, x_k, x_1)$ of the graph of $M^{(\pi)}$. Deduce that

$$\rho = \max \frac{r(x_1, x_2) + \cdots + r(x_{k-1}, x_k) + r(x_k, x_1)}{k} \ ,$$

where the maximum is taken over all cycles $(x_1, \dots, x_k, x_1)$ of $\mathcal{G}$. This scalar is called the *maximal cycle mean* of the graph $\mathcal{G}$ with weights r.

Q 4.9. Let $\beta > 0$ be a parameter and consider the nonnegative $n \times n$ matrix $A^{(\beta)}$ with entries

$$A^{(\beta)}_{xy} := \begin{cases} \exp(\beta r(x, y)) & \text{when } y \in \mathcal{C}(x) \\ 0 & \text{otherwise.} \end{cases}$$

Consider also the positive vector $V^{(\beta)} \in \mathbb{R}^{\mathcal{E}}_+$ with entries

$$V^{(\beta)}_x = \exp(\beta v_x), \quad \forall x \in \mathcal{E} \ ,$$

where $v \in \mathbb{R}^{\mathcal{E}}$ is a solution of the ergodic equation in Q 4.1. Show that

$$\exp(\beta\rho) V^{(\beta)} \leq A^{(\beta)} V^{(\beta)} \leq n \exp(\beta\rho) V^{(\beta)} \ .$$

Q 4.10. Recall that from Perron-Frobenius theorem, for any irreducible nonnegative matrix A, the spectral radius $\rho(A)$ of A is an eigenvalue associated to a positive row vector m and a positive column vector v, that is $mA = \rho(A)m$ and $Av = \rho(A)v$. Deduce that

$$\exp(\beta\rho) \leq \rho(A^{(\beta)}) \leq n \exp(\beta\rho) \ ,$$

and so

$$\lim_{\beta \to +\infty} \frac{\log \rho(A^{(\beta)})}{\beta} = \rho \ .$$

Q 4.11. Consider now the following stochastic Markov decision process (MDP) with long run time average criteria. The state space is still $\mathcal{E} = \{1, \dots, n\}$, but the action space is now $\mathcal{C}(x) \subset \mathcal{C}$ when the current state is $x \in \mathcal{E}$, where $\mathcal{C}$ is the set of subsets of $\mathcal{E}$ of cardinality 1 or 2 identified to symmetric pairs of elements of $\mathcal{E}$. Moreover, the dynamics is stochastic with X_{k+1} being equal to any element of the set U_k with uniform probability. This means that the transition probabilities are independent of time and, when u has two elements for instance, they are equal to $P(X_{k+1} = y \mid X_k = x, U_k = u) = M^{(u)}_{xy} = 1/2$ for $y \in u$ and 0 otherwise. Let $r : \mathcal{E} \times \mathcal{C} \to \mathbb{R}$ be a reward function.

We consider the maximization of a long run time average criteria as in (2), and we consider the directed graph $\mathcal{G}$, with set of nodes equal to $\mathcal{E}$ and set of arcs $\mathcal{A}$ equal to the set of $(x, y) \in \mathcal{E} \times \mathcal{E}$ such that $y \in u$ for some $u \in \mathcal{C}(x)$.

We assume again that $\mathcal{G}$ is strongly connected. Show, using results of the course, that there exists $\rho \in \mathbb{R}$ and $v \in \mathbb{R}^{\mathcal{E}}$ such that

$$\rho + v(x) = F_x(v) := \max_{u = \{y,z\} \in \mathcal{C}(x)} \left(r(x, u) + \frac{v(y) + v(z)}{2} \right) \quad \forall x \in \mathcal{E} \ ,$$

where in the condition $u = \{y, z\}$, we allow y and z to be equal (when u has cardinality 1).

Interpret ρ.

Q 4.12. Consider now the nonnegative 3-dimensional array (tensor) $A^{(\beta)}$ with entries

$$A^{(\beta)}_{xyz} := \begin{cases} \exp\left(2\beta r(x, \{y, z\})\right) & \text{when } \{y, z\} \in \mathcal{C}(x) \\ 0 & \text{otherwise.} \end{cases}$$

We say that the vector $w \in \mathbb{R}^n_+$ is a (homogeneous) Perron eigenvector of $A^{(\beta)}$ if

$$\sum_{jk=1}^{n} A^{(\beta)}_{ijk} w_j w_k = \lambda w_i^2 \quad \forall i \in \{1, \dots, n\} \ ,$$

where $\lambda \geq 0$ is what is called the spectral radius of $A^{(\beta)}$ (that is the maximum of the moduli of homogeneous eigenvalues).

Rewrite the above equation as the ergodic dynamic programming equation of a Markov decision process (MDP) with a long run time average criteria :

$$\rho + v(x) = F^{(\beta)}_x(v) \ ,$$

in which $\rho = \log(\lambda)/(2\beta)$ and $v_x = \log(w_x)/\beta$ for all $x \in \mathcal{E}$. Explain the parameters of $F^{(\beta)}$.

Q 4.13. Let $v \in \mathbb{R}^{\mathcal{E}}$ be a solution of the ergodic equation in Q 4.11.

Show that

$$\rho + v(x) \leq F^{(\beta)}_x(v) \leq \frac{\log n}{\beta} + \rho + v(x) \ .$$

Q 4.14. Denote now $\lambda^{(\beta)}$ the spectral radius of $A^{(\beta)}$. Show, using techniques of the course, that

$$\rho \leq \log \frac{\lambda^{(\beta)}}{2\beta} \leq \frac{\log n}{\beta} + \rho \ .$$