Exam of the course *Markov decision processes : dynamic programming and applications* ENSTA 5OD2A/B and M2 Optimization (Paris-Saclay University and IP Paris)

Marianne Akian

Mardi 4 novembre 2025
Durée 3h

This text contains 2 different exams :

M2 Exam consists in Problems 1, 3 and 4 it is for the students who attended and need to validate Part 1 and Part 2 of the course (to obtain a M2 grade and/or to obtain more ECTS at ENSTA). No score will be given to answers to questions of Problem 2 for these students.

ENSTA Exam consists in Problems 1 and 2 it is for the other students (ENSTA students who only need to validate Part 1 of the course (6x3 hours of ENSTA lectures). No score will be given to answers to questions of Problem 3 and 4 for these students (moreover these problems may use notions that were not teached in Part 1).

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. Problem 4 is independent of the other problems, and often questions can be solved without solving the previous questions.

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 $\overline{x}$ be the number of empty rooms at the start of the day, and let N be the number of customers that will ask for a room during the day.

We assume that N is known to the inkeeper. When a customer arrives the innkeeper proposes a price $q \in \{q_1, \dots, q_L\}$ where $0 < q_1 < \dots < q_L$. The customer will accept the offer q_i with probability p_i and refuse the offer with probability $(1 - p_i)$, where $p_1 > p_2 > \dots > p_L$. 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 $X_t \in \{0, \dots, \overline{x}\}$ the number of empty rooms between the $(t-1)$-th and the t-th arrivals of a customer, and by $U_t \in \{1, \dots, L\}$ the level of price proposed by the innkeeper. Formulate the innkeeper problem as the maximization of an expected additive payoff with finite horizon (and possibly exit time), for the Markov decision process with state process X_t, control process U_t and time horizon $T = N$. Describe the dynamics (the transition probabilities) of the MDP and the instantaneous rewards.

Q 1.2. Explain how an optimal strategy of the innkeeper can be obtained by using the following Dynamic programming equation for $t \in \{0, \dots, T-1\}$:

$$\begin{cases} v_t(x) = \max_{i \in \{1,\dots,L\}} \Big(p_i \big(q_i + v_{t+1}(x-1) \big) + (1-p_i) v_{t+1}(x) \Big), \quad x \in \{1, \dots, \overline{x}\} \\ v_t(0) = 0 \ . \end{cases}$$

Q 1.3. Show by backward induction on t, that the functions $v_t : x \in \{0, \dots, \overline{x}\} \mapsto v_t(x)$ of Q 1.2 are nondecreasing.

Q 1.4. Assume that $p_1 q_1 < \dots < p_L q_L$. Show, in that case, that the innkeeper should always charge at the highest rate q_L.

---

2 Problem 2 (to validate the ENSTA lectures only)

We consider the framework and notations of Problem 1, for which we assume now that $p_\ell q_\ell = \max\{p_i q_i, \ i = 1, \dots L\}$ for some $\ell < L$. For $t \in \{0, \dots, T-1\}$ and $x \in \{1, \dots, \overline{x}\}$, we denote

$$\iota(t,x) = \max \left(\operatorname*{Argmax}_{i \in \{1,\dots,L\}} \Big(p_i \big(q_i + v_{t+1}(x-1) \big) + (1-p_i) v_{t+1}(x) \Big) \right) \ .$$

Q 2.1. Show that $v_t \geq v_{t+1}$ for all $t \in \{0, \dots, T-1\}$.

Q 2.2. Consider the map $w_t : x \in \{1, \dots, \overline{x}\} \mapsto v_t(x) - v_t(x-1)$. Show the following properties by backward induction on $t \in \{0, \dots, T\}$:

— w_t is nonincreasing (which means that v_t is a discrete concave function);

— $x \in \{0, \dots, \overline{x}\} \mapsto v_t(x) - v_{t+1}(x)$ is nondecreasing.

Q 2.3. Show that for all $x, x' \in \{1, \dots, \overline{x}\}$ and $t \in \{0, \dots, T-1\}$, we have :

$$(p_{\iota(t,x')} - p_{\iota(t,x)})(w_{t+1}(x') - w_{t+1}(x)) \leq 0 \ .$$

Deduce that $x \in \{1, \dots, \overline{x}\} \mapsto \iota(t,x)$ is nonincreasing and give an interpretation of this property.

Q 2.4. Show that for all $x \in \{1, \dots, \overline{x}\}$, and $t.t' \in \{0, \dots, T-1\}$, we have :

$$(p_{\iota(t',x)} - p_{\iota(t,x)})(w_{t'+1}(x) - w_{t+1}(x)) \leq 0 \ .$$

Q 2.5. Deduce that, for all $x \in \{1, \dots, \overline{x}\}$, the map $t \in \{0, \dots, T-1\} \mapsto \iota(t,x)$ is nonincreasing and $\geq \ell$ and give an interpretation of these properties.

Q 2.6. 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, 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 Markov decision process with state (X_t, Y_t), in which the instananeous reward evaluates to zero when Y_t is equal to zero.

Q 2.7. Show that the value function satisfies the following Dynamic Programming equation

$$\begin{cases} v(x,1) = \max_{i\in\left\{1,\dots,L\right\}} \left(p_i\big(q_i + \alpha v(x-1,1)\big) + (1-p_i)\alpha v(x,1)\right), & x \in \{1,\dots,\overline{x}\} \\ v(0,1) = 0 & \\ v(x,0) = 0, & x \in \{0,\dots,\overline{x}\} \ . \end{cases}$$

Q 2.8. Interpret this equation as the Dynamic Programming equation of an infinite horizon discounted problem, and explain why this equation has a unique solution.

Q 2.9. How can it be solved ?

Q 2.10. Denote $\widetilde{v}(x) = v(x,1)$ and $w : x \in \{1,\dots,\overline{x}\} \mapsto w(x) = \widetilde{v}(x) - \widetilde{v}(x-1)$. Using the properties of the fixed point equation in Q 2.7, show that the map w is non increasing, meaning that $\widetilde{v}$ is concave.

Q 2.11. Deduce that the optimal policy $\pi : \{1,\dots,\overline{x}\} \to \{1,\dots,L\}$ is nonincreasing and $\geq \ell$.

Q 2.12. Assume that $L = 2$ and $\ell = 1$. Show that the number of policy iterations starting at the constant policy $\pi(x) \equiv \ell$ is at most $\overline{x}$ and compute explicitely one policy iteration.

3 Problem 3 (to validate the full (M2/Part1 and Part2) lectures only)

We consider a variant of Problem 1, in which some rooms may become unusable although they are empty, but the innkeeper observes the number of empty and usable rooms only when the customer accept the offer.

Q 3.1. Denote now by $X_t \in \{0,\dots,\overline{x}\}$ the number of empty and usable rooms between the $(t-1)$-th and the t-th arrivals of a customer, by $U_t \in \{1,\dots,L\}$ the level of price proposed by the innkeeper, by $C_{t+1} \in \{0,1\}$ the decision of the customer after this period of time : $C_{t+1} = 1$ if he accepts the offer and $C_{t+1} = 0$ otherwise, and by Y_{t+1} the observation after this period : $Y_{t+1} = X_{t+1}$ if $C_{t+1} = 1$ and $Y_{t+1} = -1$ (meaning no observation) otherwise. We also assume that the number of empty and usable rooms evolves as follows : $X_{t+1} = \max(X_t - C_{t+1} - W_t, 0)$ with W_t a sequence i.i.d random variables with values in $\{0,\dots,\overline{x}\}$ and law $\varphi : \varphi(w) = \mathbb{P}\left(W_t = w\right)$. The probabilities of acceptation of customers are the same as in Problem 1. No penalty occurs when the offer is accepted but no room is available, that is if $X_t - C_{t+1} - W_t < 0$. In that case the reward of the innkeeper is zero.

Formulate the innkeeper problem as a Partially Observable Markov Decision Process (POMDP) with state process (X_t, C_t), control process U_t, observation process Y_t and additive payoff with finite time horizon $T = N$. Describe the transition probabilities of the POMDP and the instantaneous rewards.

Q 3.2. Compute the dynamics of the belief and show that after one step of the POMDP, starting at some belief b, one can only reach the Dirac measures (in any point of $\{0,\dots,\overline{x}\}$) and bM where M is the transition matrix of X_t when there are no customers : $M_{xx'} = \mathbb{P}\left(\max(x - W_t, 0) = x'\right)$.

Q 3.3. Write the dynamic programming equation satisfied by the value $v_t(b,c)$ of the POMDP as a function of the belief b at step t on the state variable x, and of the decision c of the customer. Show in particular that v_t does not depend on c.

4 Problem 4 (to validate the full (M2/Part1 and Part2) 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 equal to $\mathbb{P}\left(X_{k+1} = y \mid X_k = x, U_k = u\right) = M_{xy}^{(u)} = \delta_{yu}$ (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 criterion, among all state and control processes $(X_k, U_k)_{k \geq 0}$ determined by any strategy σ and starting at some state $x \in \mathcal{E}$:

$$J^\sigma(x) = \limsup_{T \to \infty} \left\{ \frac{1}{T} \mathbb{E}^\sigma \left[\sum_{k=0}^{T} r(X_k, U_k) \mid X_0 = x \right] \right\} \quad , \tag{1}$$

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

We associate to the above MDP 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

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

and relate the solution with the value of $\zeta(x)$, for $x \in \mathcal{E}$.

Q 4.2. Let π be a policy, that is an element of $\Pi := \{\pi : \mathcal{E} \to \mathcal{C} \mid \pi(x) \in \mathcal{C}(x), \ \forall x \in \mathcal{E}\}$ (the set of stationary pure Markov strategies), and consider (following the notations of the course) the vector and matrix

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

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 4.3. 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 4.4. 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 4.5. Recall that $\mathcal{B}(v) \geq r^{(\pi)} + M^{(\pi)} v$ for all $v \in \mathbb{R}^{\mathcal{E}}$ and $\pi \in \Pi$. Deduce that

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

Q 4.6. Show that indeed the previous inequality holds for all cycles $(x_1, \dots, x_k, x_1)$ of $\mathcal{G}$.

Q 4.7. Let π be an optimal policy for a solution v of the ergodic equation in Q 4.1, that is a policy such that

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

Show that

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

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

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

Let $w^{(\beta)}$ be a positive eigenvector of $A^{(\beta)}$ associated to its spectral radius $\lambda^{(\beta)} = \rho(A^{(\beta)})$, meaning :

$$A^{(\beta)} w^{(\beta)} = \lambda^{(\beta)} \lambda w^{(\beta)} \ ,$$

and such that $w^{(\beta)} \mathbf{1} = 1$. Such a vector $w^{(\beta)}$ exists and is unique by Perron-Frobenius theorem.

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

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

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

Q 4.9. Let (ρ, v), with $\rho \in \mathbb{R}$ and $v \in \mathbb{R}^{\mathcal{E}}$, be a solution of the ergodic equation in Q 4.1.

Show that

$$\rho + v(x) \leq [\mathcal{B}^{(\beta)}(v)](x) \leq \frac{\log n}{\beta} + \rho + v(x) \quad \forall x \in \mathcal{E} \ .$$

Q 4.10. Deduce, using techniques of the course, that

$$\rho \leq \frac{\log(\lambda^{(\beta)})}{\beta} \leq \frac{\log n}{\beta} + \rho \quad \text{so} \quad \lim_{\beta \to +\infty} \frac{\log(\lambda^{(\beta)})}{\beta} = \rho \ .$$

Q 4.11. Assume that the solution v of Q 4.1 is unique up to an additive constant, and let $w^{(\beta)}$ be as above. Show that $w_x = \lim_{\beta \to +\infty} \frac{\log(w^{(\beta)}_x)}{\beta}$ exists for all $x \in \mathcal{E}$ and that $w = (w_x)_{x \in \mathcal{E}}$ is a solution of the ergodic equation in Q 4.1 : $\rho \mathbf{1} + w = \mathcal{B}(w)$.