Résolution numérique de problèmes de commande optimale de chaînes de Markov observées imparfaitement

Nadir Farhi, Jean-Pierre Quadrat

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Résolution numérique de problèmes de commande optimale de chaînes de Markov observées imparfaitement

Nadir Farhi — Jean-Pierre Quadrat

No 5348

October 2004

THÈME 4

ISSN 0249-6399

Résolution numérique de problèmes de commande optimale de chaînes de Markov observées imparfaitement

Nadir Farhi* , Jean-Pierre Quadrat*

Thème 4 — Simulation et optimisation
de systèmes complexes
Projet Metalau

Rapport de recherche no 4 — October 2004 — 19 pages

Résumé : Nous résolvons numériquement le problème de commande optimale de chaînes de Markov en observations incomplètes. Grâce à une quantification des filtres optimaux possibles on approxime le support de la loi du filtre optimal. On résout l'équation de la programmation sur cet espace d'états quantifiés. Sur deux exemples on montre l'effectivité de ce point de vue : – un problème d'embauche d'une secrétaire, – une problème de renouvellement optimal

Mots-clés : Contrôle optimal stochastique, Filtrage optimal, Problème de l'embauche d'une secrétaire, Problème de renouvellement

* INRIA-Rocquencourt

Numerical Solution of Stochastic Control Problem with Incomplete Observations

Abstract: We solve numerically optimal control problem of finite Markov chains in the case of incomplete observations. By quantifying the possible optimal filters we obtain an approximation of the support of the probability law of the optimal filter. We solve the dynamic programming equation on this space of quantified states. On two examples we show that this point of view is effective : – a secretary job problem, – a standard optimal renewal problem.

Key-words: Optimal Stochastic control, optimal filtering, Secretary Problem, Optimal Renewal Rroblem, Reliability

1 Introduction

Le but de ce travail est la résolution numérique de l'équation de la programmation dynamique (EPD) pour des problèmes de commande optimale stochastique de chaînes de Markov incomplètement observées. L'EPD de ce problème est donnée dans [1]. On la trouve maintenant dans des cours standards [11]. Dans le cas d'observation incomplète, l'état permettant d'écrire la récurrence de la programmation dynamique est la probabilité conditionnelle de l'état de la chaîne de Markov connaissant le passé des observations (appelée filtre optimal). Cette loi conditionnelle est une chaîne de Markov vivant dans un simplexe de dimension égal au nombre d'états de la chaîne de Markov (supposé fini dans tout ce travail). La résolution de l'EPD est insoluble numériquement dans toute sa généralité dès que la chaîne de Markov possède plus que 5 ou 6 états. On trouvera l'état de l'art pour la résolution numérique de ce problème dans [6].

Cependant dans certains cas particuliers le support de la loi du filtre peut être approximé par un nombre fini raisonnable de points du simplexe. Dans ces cas particuliers on peut ramener le problème initial à un problème de commande de chaînes de Markov en observation complète avec un nombre raisonnable d'états. L'approximation du support est obtenue par quantification du filtre optimal à chaque étape de son calcul récursif. Ce point de vue semble original bien que plusieurs méthodes voisines aient été proposées (voir [6]).

Nous illustrons ce propos sur deux exemples : – le problème du recrutement d'une secrétaire qui a été abondamment étudié par exemple [2, 3, 4, 5, 7, 8, 12] – un problème de remplacement optimal classique dont une référence récente est [9]. Sur ces deux exemples nous construisons la chaîne de Markov approximant le filtre optimal puis nous résolvons l'EPD numériquement par l'algorithme de Howard disponible dans la boîte à outils Maxplus de Scilab [13].

2 Rappels

2.1 Chaîne de Markov observée

Une chaîne de Markov observée est définie à partir du 5-uple : $(\mathcal{T}, \mathcal{E}, \mathcal{G}, M^y, P^y)$, où l'on appelle :

- $\mathcal{T}$: l'espace des temps qui sera ici $\mathbb{N}$,
- $\mathcal{E}$: l'ensemble fini des états,
- $\mathcal{G}$: l'ensemble fini des observations,
- $M^y_{xx'}$: la probabilité d'aller en $x' \in \mathcal{E}$ et d'observer $y \in \mathcal{G}$ partant de $x \in \mathcal{E}$,
- P^y_x : la probabilité d'être en $x \in \mathcal{E}$ et d'observer $y \in \mathcal{G}$ à l'instant initial.

Nous avons alors :

$$\sum_{y\in\mathcal{G}} \sum_{x'\in\mathcal{E}} M^y_{xx'} = 1 \ , \forall x \in \mathcal{E}$$

$$\sum_{y\in\mathcal{G}} \sum_{x\in\mathcal{E}} P^y_x = 1 \ .$$

L'ensemble des trajectoires des couples états-observations $\Omega = (\mathcal{E} \times \mathcal{G})^{\mathcal{T}}$ est muni d'une loi de probabilité markovienne définie de façon unique par la probabilité des cylindres de trajectoires :

$$\mathbb{P}((x_0, y_0), (x_1, y_1), ..., (x_n, y_n)) = P_{x_0}^{y_0} M_{x_0 x_1}^{y_1} M_{x_1 x_2}^{y_2} ... M_{x_{n-1} x_n}^{y_n} \ .$$

Pour $\omega = (\omega^1 \in \mathcal{E}^{\mathcal{T}}, \omega^2 \in \mathcal{G}^{\mathcal{T}}) \in \Omega$ on note $X_n(\omega) = \omega_n^1 \in \mathcal{E}$ l'état de la chaîne de Markov et $Y_n(\omega) = \omega_n^2 \in \mathcal{G}$ l'observation à l'instant $n \in \mathcal{T}$.

A chaque instant n on dispose des observations $Y(0), \cdots, Y(n)$. Lorsque $\mathcal{E} = \mathcal{G}$ et $M_{xx'}^y = N_{xx'} I_{x'y}$ (où I désigne la matrice identité) et $P_x^y = R_x I_{xy}$ on dit que la chaîne de Markov est observée complètement. On peut alors simplifier la définition du processus en ne conservant que ce qui concerne l'état dans les définitions précédentes ($\Omega = \mathcal{E}^{\mathcal{T}}$, $\mathbb{P}((x_0, \cdots x_n)) = R_{x_0} N_{x_0 x_1} \cdots N_{x_{n-1} x_n}$).

Le problème fondamental des chaînes de Markov observées est la détermination de l'équation récursive du filtre optimal donnant la probabilité conditionnelle de l'état connaissant le passé des observations :

$$Q^n = \mathbb{P}(X_n \mid Y_0, \cdots, Y_n)$$

qui vérifie :

$$Q^{n+1} = \frac{Q^n M^{Y_n}}{Q^n M^{Y_n} \mathbf{1}}, \quad Q^0 = \frac{P^{Y_0}}{P^{Y_0} \mathbf{1}} \ ,$$

avec $\mathbf{1} = (1, \cdots, 1)'$. Le processus $Q_n(\omega)$ étant une loi de probabilité sur $\mathcal{E}$ vit dans le simplexe $\mathcal{S}_E = \{q \mid \sum_{x \in \mathcal{E}} q_x = 1, \ q_x \geq 0\}$. Ce processus est markovien et sa loi est définie par le noyau de transition :

$$\mathbf{M}_{qq'} = \begin{cases} qM^y \mathbf{1} & \text{si } q' = \frac{qM^y}{qM^y \mathbf{1}}, \ y \in \mathcal{G} \ , \\ 0 & \text{sinon.} \end{cases}$$

Les supports des lois invariantes du filtre optimal sont mal connus. Ils peuvent être denses dans le simplexe ou sur un sous ensemble du simplexe ou bien être ponctuels. Nous explorerons ce support sur des exemples.

2.2 Chaîne de Markov commandée temps invariante

Une chaîne de Markov commandable est définie à partir du 6-uple : $(\mathcal{T}, \mathcal{E}, \mathcal{F}, M^u, P, c^u)$, où l'on appelle :

- $\mathcal{T}$: l'espace des temps qui sera ici $\mathbb{N}$,
- $\mathcal{E}$: l'ensemble fini des états,
- $\mathcal{F}$: l'ensemble fini des commandes,
- $M_{xx'}^u$: la probabilité d'aller en $x' \in \mathcal{E}$ sachant que la commande choisie est $u \in \mathcal{F}$ et que l'état de départ est $x \in \mathcal{E}$,
- P_x : la probabilité d'être en $x \in \mathcal{E}$ à l'instant initial,

– c_x^u le coût sur une période de temps d'avoir pris la décision $u \in \mathcal{F}$ et d'être dans l'état $x \in \mathcal{E}$.

Nous avons alors :

$$\sum_{x' \in \mathcal{E}} M_{xx'}^u = 1 \ , \forall x \in \mathcal{E}, \ \forall u \in \mathcal{F} \ ,$$

$$\sum_{x \in \mathcal{E}} P_x = 1 \ .$$

Pour pouvoir déterminer la loi de probabilité d'une chaîne de Markov commandée il faut préciser la façon de choisir les commandes. On le fait ici en se donnant une stratégie stationnaire. Une stratégie stationnaire est la donnée d'une application $s : \mathcal{E} \mapsto \mathcal{F}$ (l'ensemble des stratégies est noté $\mathcal{S}$).

Etant donnée une chaîne de Markov commandable et une stratégie stationnaire, on définit sur l'ensemble des trajectoires des états $\Omega = \mathcal{E}^{\mathcal{T}}$ une loi de probabilité markovienne, de façon unique, par les probabilité des cylindres de trajectoires :

$$\mathbb{P}((x_0, x_1, ..., x_n)) = P_{x_0} M_{x_0 x_1}^{s(x_0)} M_{x_1 x_2}^{s(x_1)} ... M_{x_{n-1} x_n}^{s(x_{n-1})} \ .$$

Le problème fondamental des chaînes de Markov commandées est le calcul de la stratégie minimisant – soit la fonctionnelle additive de la trajectoire $\mathbb{E}J(s)$ définie par

$$J(s) = \sum_{k=0}^{\infty} c_{X_k}^{s(X_k)} \ ,$$

lorsque cette somme est finie pour au moins une stratégie s, – soit lorsque cette somme est infinie un critère fini du type :

$$J(s) = \lim_{\lambda \to 0} \lambda \sum_{k=0}^{\infty} \frac{c_{X_k}^{s(X_k)}}{(1+\lambda)^{(k+1)}} \ .$$

Pour résoudre ce problème on introduit la fonction valeur

$$v : x \in \mathcal{E} \mapsto v_x = \min_{s \in \mathcal{S}} \mathbb{E}\{J(s) \mid X_0 = x\} \ .$$

Cette fonction valeur vérifie dans le premier cas l'équation de la programmation dynamique :

$$v_x = \min_{u \in \mathcal{F}} \{(M^u v + c^u)_x\} \ . \tag{1}$$

Dans le deuxième cas, sous l'hypothèse, par exemple, que quelque soit la stratégie s la chaîne de Markov commandée par s est ergodique,on peut montrer que $v_x = v \in \mathbb{R}$ pour tout x et qu'il existe $w : x \in \mathcal{E} \mapsto w_x \in \mathbb{R}$ vérifiant :

$$w_x + v = \min_{u \in \mathcal{F}} \{(M^u w + c^u)_x\} \ . \tag{2}$$

La fonction w est alors définie à une constante additive près.

Les problèmes de temps d'arrêt, pour lesquels la fonctionnelle $J(s)$ devient

$$J(s,\nu) = \sum_{k=0}^{\nu} c_{X_k}^{s(X_k)} + \Phi_{X_\nu} \ ,$$

où ν est un temps d'arrêt et $\Phi : \mathcal{E} \mapsto \mathbb{R}$ un coût final, se ramène au cas précédent par :

- l'adjonction d'un état supplémentaire : ϕ,
- l'adjonction d'une commande supplémentaire : a,
- la définition d'un nouveau coût instantané :

$$c'^u_x = \begin{cases} c^u_x & \text{si } x \in \mathcal{E}, u \in \mathcal{F} \ , \\ \Phi_x & \text{si } x \in \mathcal{E}, u = a \ , \\ 0 & \text{si } x = \phi \ , \end{cases}$$

- la définition de nouvelles matrices de transitions :

$$M'^u_{xx'} = \begin{cases} M^u_{xx'} & \text{si } x, x' \in \mathcal{E}, u \in \mathcal{F} \ , \\ 1 & \text{si } x \in \mathcal{E}, x' = \phi, u = a \ , \\ 1 & \text{si } x = x' = \phi \ , \\ 0 & \text{sinon.} \end{cases}$$

En utilisant la structure particulière des données, et en supposant qu'il existe une stratégie pour laquelle l'arrêt se produit en temps fini presque sûrement, l'équation de la programmation dynamique peut aussi s'écrire :

$$w_x = \left[\min \left\{ \Phi, \min_{u \in \mathcal{F}} \{ M^u w + c^u \} \right\} \right]_x \ . \tag{3}$$

2.3 Chaîne de Markov commandée observée temps invariante

Une chaîne de Markov commandable est définie à partir du 7-uple :

$$(\mathcal{T}, \mathcal{E}, \mathcal{F}, \mathcal{G}, M^{uy}, P^y, c^u) \ ,$$

où l'on appelle :

- $\mathcal{T}$: l'espace des temps qui sera ici $\mathbb{N}$,
- $\mathcal{E}$: l'ensemble fini des états de cardinalité E.
- $\mathcal{F}$: l'ensemble fini des commandes,
- $\mathcal{G}$: l'ensemble fini des observations,
- $M^{uy}_{xx'}$: la probabilité d'aller en $x' \in \mathcal{E}$ et d'observer $y \in \mathcal{G}$ sachant que la commande choisie est $u \in \mathcal{F}$ et que l'état de départ est $x \in \mathcal{E}$,
- P^y_x : la probabilité d'être en $x \in \mathcal{E}$ et d'observer $y \in \mathcal{G}$ à l'instant initial,

– c_x^u le coût sur une période de temps d'avoir pris la décision $u \in \mathcal{F}$ et d'être dans l'état $x \in \mathcal{E}$.

Nous avons alors :

$$\sum_{x' \in \mathcal{E}, y \in \mathcal{G}} M_{xx'}^{uy} = 1 \ , \forall x \in \mathcal{E}, \ \forall u \in \mathcal{F} \ ,$$

$$\sum_{x \in \mathcal{E}, y \in \mathcal{G}} P_x^y = 1 \ .$$

Pour pouvoir déterminer la loi de probabilité d'une chaîne de Markov commandée il faut préciser la façon de choisir les commandes. Une commande à l'instant n ne peut dépendre que des observations passées. On le fait ici en se donnant une stratégie stationnaire sur le filtre optimal qui est un résumé exhaustif de l'information concernant l'état contenue dans les observations. Une stratégie stationnaire est alors définie au moyen d'une application $z : \mathcal{S}_E \mapsto \mathcal{F}$ où $\mathcal{S}_E$ désigne l'ensemble des probabilités sur $\mathcal{E}$ (l'ensemble des stratégies est noté $\mathcal{Z}$). Au moyen d'une stratégie z on définit des fonctions du temps $(Z_n)_{n \in \mathcal{T}}$ dépendant des seules observations passées par composition avec la loi du filtre optimale $Z_n(y_0, \cdots, y_n) = z \circ q^n$ où q^n est le filtre optimal basé sur les observations $y_0, \cdots, y_n$ associée à la chaîne de Markov commandée par z jusqu'à l'instant n. On peut vérifier par récurrence que l'on a par ce procédé défini de façon unique les fonctions Z_n.

Etant donnée une chaîne de Markov commandable et une stratégie z stationnaire on définit sur l'ensemble des trajectoires des états-observations $\Omega = (\mathcal{E} \times \mathcal{G})^{\mathcal{T}}$ une loi de probabilité markovienne, de façon unique, par les probabilité des cylindres de trajectoires :

$$\mathbb{P}(\{(x_0, y_0), (x_1, y_1), ..., (x_n, y_n)\}) = P_{x_0}^{y_0} M_{x_0 x_1}^{Z_0(y_0), y_1} ... M_{x_{n-1} x_n}^{Z_{n-1}(y_0, \cdots, y_{n-1}), y_n} \ .$$

Le problème fondamental des chaînes de Markov commandées est le calcul de la stratégie minimisant la fonctionnelle additive de la trajectoire $\mathbb{E}J(z)$ définie par

$$J(z) = \sum_{k=0}^{\infty} c_{X_k}^{Z_k(Y_0, \cdots, Y_k)} \ ,$$

lorsque cette somme est finie[1] pour au moins une stratégie z. Pour résoudre ce problème on introduit la fonction valeur non plus sur l'état mais sur l'ensemble des probabilités sur l'espace des états et on se ramène à un coût sur le filtre optimal

$$v : q \in \mathcal{S}_E \mapsto v(q) = \min_{z \in \mathcal{Z}} \mathbb{E}\{J(z) \mid Q_0 = q\} \ .$$

[1] Lorsque cette somme est infinie on utilise un critère fini du type :

$$J(z) = \lim_{\lambda \to 0} \lambda \sum_{k=0}^{\infty} \frac{c_{X_k}^{Z_k(Y_0, \cdots, Y_k)}}{(1+\lambda)^{(k+1)}} \ .$$

En effet en utilisant les propriété de l'espérance conditionnelle sur des tribus emboîtées on a par exemple pour le premier critère

$$\mathbb{E}J(z) = \mathbb{E}\sum_{0}^{\infty} Q^k.c^{z(Q^k)},$$

où l'opérateur binaire "." désigne le produit scalaire de $\mathbb{R}^E$ et où les trajectoires du filtre vérifient :

$$Q^{k+1} = \frac{Q^k M^{z(Q_k),Y^k}}{Q^k M^{z(Q_k),Y^k}\mathbf{1}}.$$

Le processus filtre optimal est la chaîne de Markov, commandée par z, temps invariante, définie par le quintuplet $(\mathcal{T}, \mathcal{S}_E, \mathcal{F}, \mathbf{M}^u, q.c^u)$ où :

$$\mathbf{M}^u_{qq'} = \begin{cases} qM^{uy}\mathbf{1} & \text{si } q' = \frac{qM^{uy}}{qM^{uy}\mathbf{1}},\ u \in \mathcal{F},\ y \in \mathcal{G}\ , \\ 0 & \text{sinon.} \end{cases}$$

Cette fonction valeur vérifie l'équation de la programmation dynamique :

$$v(q) = \min_{u\in\mathcal{F}} \left\{ \sum_y \left\{ v\left(\frac{qM^{uy}}{qM^{uy}\mathbf{1}}\right) qM^{uy}\mathbf{1} \right\} + q.c^u \right\}\ . \tag{4}$$

Dans la suite nous montrons, sur deux exemples, comment on peut résoudre numériquement cette équation bien que l'on soit en présence d'un problème en dimension grande (à priori insoluble par les méthodes standard de l'analyse numérique consistant à discrétiser l'espace $\mathcal{S}_E$).

3 Problème de l'embauche d'une secrétaire

Nous résolvons numériquement ici une variante du problème de l'embauche d'une secrétaire. Nous avons choisi une variante conduisant à un problème de chaîne de Markov temps invariante.

Nous voulons embaucher une secrétaire. Il y a E candidates pour le poste, toutes de valeurs différentes. On suppose que les valeurs possibles des candidates sont $\mathcal{E} = \{1, \cdots, E\}$ car on veut seulement les comparer (et en fait on ne s'intéresse qu'à la meilleure qui est celle de valeur 1). Pour embaucher la secrétaire on fait des interviews. A un moment, qu'il faut optimiser, on arrête les interviews et on en embauche la meilleure secrétaire rencontrée jusque là. On peut interviewer plusieurs fois la même secrétaire. L'interviewée est tirée au hasard avec remise parmi les E secrétaires. Lorsqu'une secrétaire a été interviewée on observe, non pas son niveau, mais seulement, le fait qu'elle soit meilleure ou non que toutes celles interviewées jusque là. On veut minimiser un coût qui est un compromis entre le temps passé pour réaliser cette embauche et la probabilité, que l'on aimerait grande, de trouver la meilleure secrétaire.

3.1 Modélisation en terme de temps d'arrêt optimal

Le problème de l'embauche d'une secrétaire est le problème de temps d'arrêt optimal de la chaîne de Markov observée définie par le 7-uple suivant :

$$(\mathcal{T}, \mathcal{E}, \mathcal{G}, M^y, P^y, c, \Phi)$$

avec :

1. $\mathcal{T} = \mathbb{N}$.
2. $\mathcal{E} = \{1, \cdots, E\}$ l'ensemble des états. L'état de ce système $X^k \in \mathcal{E}$ est la valeur de la meilleure secrétaire rencontrée jusqu'à l'instant k. Cet état n'est pas observée.
3. $\mathcal{G} = \{0, 1\}$ l'ensemble des observations. L'observation $Y^k = 1$ signifie que la secrétaire observée à l'instant k est meilleure que toutes celles observées jusqu'à l'instant $k-1$, $Y^k = 0$ signifie le contraire la secrétaire observée à l'instant k a une valeur inférieure ou égale à celles observées jusqu'à l'instant $k-1$.
4. La secrétaire interviewée étant tirée avec la loi uniforme avec remise sur l'ensemble des secrétaires la matrice de transition de la chaîne de Markov représentant la valeur de la meilleure secrétaire observée est

$$M = \begin{bmatrix} 1 & 0 & \cdot & \cdot & 0 \\ 1/E & 1-1/E & 0 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 1/E & 1/E & \cdot & \cdot & 1/E \end{bmatrix} .$$

   Alors on a :

$$M^0 = \mathrm{diag}(M), \quad M^1_{xx'} = \begin{cases} M_{xx'} & \text{si } x < x',\ x, x' \in \mathcal{E} , \\ 0 & \text{sinon.} \end{cases}$$

   En effet $M^0_{xx'}$ [resp. $M^1_{xx'}$] signifie probabilité que la meilleure secrétaire interviewée soit de valeur x' sachant qu'on n'a pas observé d'amélioration de valeur [resp. la dernière observée est meilleure] et que le niveau de la meilleure secrétaire observée avant la dernière interview est x.
5. La loi initiale est donnée par

$$P^1_x = 1/E, \quad P^0_x = 0, \quad \forall x \in \mathcal{E} ,$$

   puisque on ne dispose à priori d'aucune information sur les secrétaires.
6. Le coût instantané d'une interview est pris égal à μ/E ($\mu \in \mathbb{R}$ donné) de façon à ce que le nombre d'interviews optimal soit de l'ordre du nombre de secrétaires :

$$c_x = \mu/E,\ \forall x \in \mathcal{E} .$$

7. Le coût d'arrêt est
$$\Phi_x = \begin{cases} 1 & \text{si } x = -1 \ , \\ 0 & \text{sinon} \ , \end{cases}$$
en effet ce coût final conditionné par le passé des observations donne l'opposé de la probabilité de choisir la meilleure secrétaire.

L'équation de programmation s'écrit alors :
$$v(q) = \min \left\{ -q_1, \mu/E + \sum_{y \in \{0,1\}} v\left(\frac{qM^y}{qM^y\mathbf{1}}\right) qM^y\mathbf{1} \right\}, \ \forall q \in \mathcal{S}_E \ . \tag{5}$$

3.2 Résolution Numérique

La principale difficulté dans la résolution numérique de (6) provient du caractère infini de $\mathcal{S}_E$ dont la discrétisation nécessite une mémoire de taille croissant de façon exponentielle avec le nombre de secrétaires E. Au lieu de discrétiser $\mathcal{S}_E$ nous allons approximer le support loi du filtre Q^k. Pour cela nous considérons l'opérateur de quantification $\mathbb{Q}$ conservant les l premiers bits des composantes d'un point de $\mathcal{S}_E$ plus précisément, en notant $r = 2^l - 1$ on définit $\mathbb{Q}$ par :
$$\mathbb{Q}_l : q \in \mathcal{S}_E \mapsto \frac{1}{r}\left(\lfloor rq_1 \rfloor, \cdots, \lfloor rq_{E-1} \rfloor, r - \sum_{i=1}^{E-1} \lfloor rq_i \rfloor \right) \in \mathcal{S}_E \cap (\mathbb{N}/r)^E \ .$$

Avec cette définition $\mathbb{Q}(q)$ est une loi de probabilité. On prolonge le domaine de définition de $\mathbb{Q}_l$ à $\mathbb{R}^E$ par $\bar{\mathbb{Q}}_l(q) = \mathbb{Q}(q/q.\mathbf{1})$ pour $q \in \mathbb{R}^E$.

On considère alors la solution stationnaire $\mathcal{D} \subset \mathcal{S}_E$ de :
$$\mathcal{D}^k = \bar{\mathbb{Q}}_l(\mathcal{D}^{k-1} \cup \mathcal{D}^{k-1}M^0 \cup \mathcal{D}^{k-1}M^1), \quad \mathcal{D}^0 = \{P^1\} \ ,$$
équation récurrente qui se stationnarise après un nombre fini d'itérations du fait de la quantification qui borne la cardinalité des ensembles $\mathcal{D}^k$. Cet ensemble fini approxime le support de la loi du filtre optimal.

On approxime alors l'équation (6) par une équation de la programmation dynamique définie sur $\mathcal{D}$:
$$v_l(d) = \min \left\{ -d_1, \mu/E + \sum_{y \in \{0,1\}} v_l\left(\mathbb{Q}_l\left(\frac{dM^y}{dM^y\mathbf{1}}\right)\right) dM^y\mathbf{1} \right\}, \ \forall d \in \mathcal{D} \ . \tag{6}$$

Cette équation de la programmation dynamique est résolue par l'algorithme de Howard [11] implémenté dans la boîte à outils Maxplus de Scilab [13].

Les deux tableaux qui suivent montrent que la complexité de ce problème reste raisonnable en taille mémoire et en temps de calcul. Ainsi le problème avec 200 secrétaires a été résolu sur une machine ancienne ayant peu de mémoire centrale.

	Niveau de quantification l							
E	3	4	5	6	7	8	9	10
3	4	9	13	17	24	29	37	43
5	6	12	19	35	54	104	161	252
10	×	11	23	45	83	198	458	1113
20	×	×	20	49	91	193	528	1468
30	×	×	17	42	92	207	472	1455
50	×	×	×	38	119	205	429	1161
100	×	×	×	×	51	154	413	870
200	×	×	×	×	×	107	290	983

TAB. 1 – Nombre d'états quantifiés (cardinal de $\mathcal{D}$) en fonction du nombre de secrétaires E et du nombre de bits l utilisé pour la quantification.

	Niveau de quantification l							
E	3	4	5	6	7	8	9	10
3	00.06	00.10	00.12	00.09	00.12	00.13	00.14	00.17
5	00.07	00.11	00.10	00.16	00.18	00.26	00.40	00.63
10	×	00.09	00.18	00.20	00.29	00.54	01.62	06.99
20	×	×	00.21	00.32	00.40	00.72	02.04	09.58
30	×	×	00.22	00.27	00.49	00.95	02.39	11.39
50	×	×	×	00.37	01.10	01.59	03.34	12.70
100	×	×	×	×	00.82	02.03	09.27	21.49
200	×	×	×	×	×	02.99	10.84	79.93

TAB. 2 – Temps de calcul, en seconde (PC 400Mhz), pour la résolution du problème d'embauche de la secrétaire en fonction du nombre de secrétaire E et du nombre de bits l utilisé pour la quantification.

Pour représenter la stratégie optimale il suffit d'indiquer les états dans lesquels on arrête les interviews. Ces états seront représentés par des gros cercles rouges dans les deux diagrammes suivants (les autres états par des points ou des petits cercles verts). Pour représenter dans le plan ou sur une droite des points de $\mathcal{D}$ (appartenant à un espace de dimension E) nous avons utilisé le codage suivant :

$$\mathbb{C} : d \in \mathcal{D} \mapsto ((d_1 2^l) + d_2)2^l + \cdots + d_E \in \mathbb{R} .$$

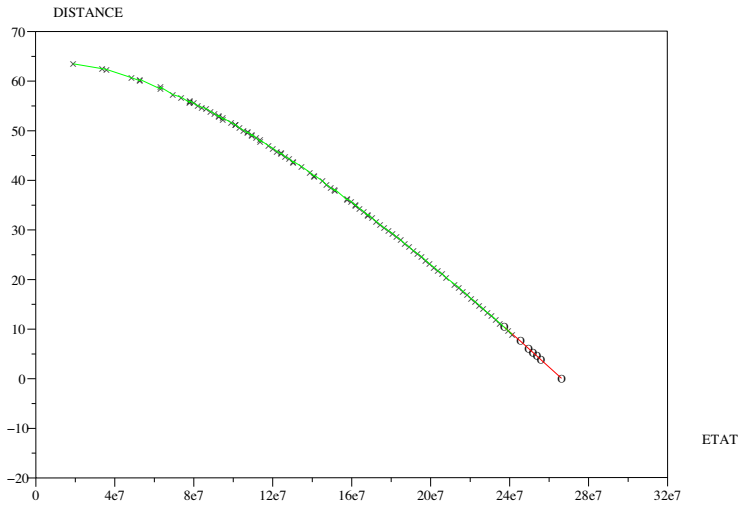

FIG. 1 – Etats de la chaîne de Markov sur $\mathcal{D}$ et stratégie optimale dans le cas : – 20 secrétaires, – quantification à 8 bits, – représentation $d \mapsto (\mathbb{C}(d), \mathbb{D}(d))$.

Le filtre optimal converge presque sûrement vers $e_1 = (1, 0, \cdots, 0)$ (probabilité de 1 sur la meilleure secrétaire) et donc la chaîne de Markov induite sur $\mathcal{D}$ a une seule classe finale (le point e_1) dans laquelle rentre, presque sûrement en temps fini, toutes les trajectoires de la chaîne de Markov sur $\mathcal{D}$. Les états dans lesquels sont arrêtés les interviews doivent être dans voisinage de cet état. Pour montrer cela on introduit la distance d'un point de $d \in \mathcal{D}$ à e_1 définie par

$$\mathbb{D} : d \in \mathcal{D} \mapsto \mathbb{D}(d) = (1 - d_1)(E - 1) + \sum_{i=2}^{E} (E - i)d_i \in \mathbb{R} .$$

On voit sur la figure 1 que la stratégie optimale s'exprime quasiment en fonction de cette seule distance. On arrête les interviews dès que $\mathbb{D}(d)$ est plus petite qu'un certain seuil.

Pour être plus précis on peut représenter dans un plan les états de $\mathcal{D}$ en les classant au moyen de deux paramètres :

1. *L'abscisse* : représentant l'ordre induit sur $\mathcal{D}$ par $\mathbb{R}$ par le codage $\mathbb{C}$ ($d^1 \geq d^2$ si $\mathbb{C}(d^1) \geq \mathbb{C}(d^2)$).
2. *L'ordonnée* : le nombre d'améliorations de la valeur des secrétaires observées pour obtenir l'état d. En effet un point d est en correspondance avec les mots sur l'alphabet $\{0, 1\}$ indiquant les observations qu'il a fallu faire pour atteindre cet état. Les mots possibles sont de longueurs arbitraires mais contiennent au plus E fois le nombre 1. L'ordonnée donne le nombre maximum de 1 des mots conduisant à l'état considéré.

La figure 2 représente la chaîne de Markov sur $\mathcal{D}$ dans ce plan (que nous appellerons quantification-amélioration).

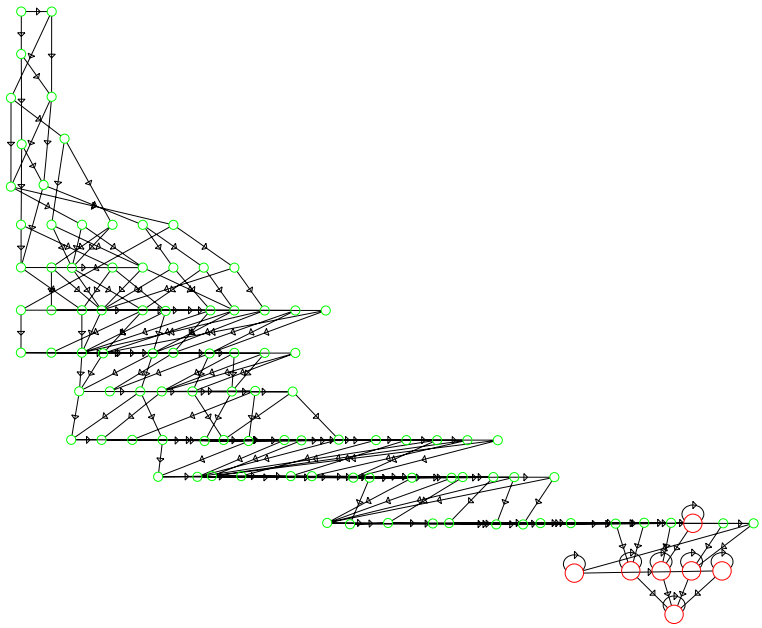

FIG. 2 – Représentation dans le plan quantification-amélioration de la chaîne de Markov sur $\mathcal{D}$ correspondant au problème avec 15 secrétaires et un codage sur 8 bits.

4 Problème de maintenance

Nous disposons d'un appareil dont l'état de dégradation est représenté par un entier $x \in \mathcal{E} = \{1, ..., E\}$. Le nombre 1 correspond à l'état neuf et le nombre E à l'état de panne. On se donne la probabilité λ que l'appareil se dégrade d'un niveau en une période de temps. Nous supposons que pour connaître l'état de l'appareil il faut faire un test de coût τ. A chaque instant on doit choisir une décision $u \in \mathcal{F} = \{0, 1, 2\}$ où : – 0 signifie on ne fait rien, – 1 signifie on teste l'appareil, – 2 on remplace l'appareil qui coûte ρ. On suppose pour simplifier la formulation du problème que l'on ne peut pas tester et remplacer l'appareil pendant la même période de temps. On se donne, de plus, un coût d'utilisation par période π_x dépendant de l'état de l'appareil x (par exemple une voiture dont la carburation est mal réglée a une consommation dépendant de l'état de la carburation). On veut optimiser la politique d'entretien de l'appareil.

4.1 Modélisation en terme de commande optimale stochastique

Le problème de maintenance est le problème de commande optimal de chaîne de Markov défini par le 7-uple suivant :

$$(\mathcal{T}, \mathcal{E}, \mathcal{F}, \mathcal{G}, M^{uy}, P^y, c^u)$$

avec :

1. $\mathcal{T} = \mathbb{N}$.
2. $\mathcal{E}$ les états de dégradations de l'appareil.
3. $\mathcal{F}$ les décisions possibles de maintenance.
4. $\mathcal{G} = \{0\} \cup \mathcal{E}$ les observations possibles, 0 signifie pas d'observation, $y \in \mathcal{E}$ signifie que l'on a observé (grâce au test) l'état de dégradation de l'appareil,
5. En notant :
$$M = \begin{bmatrix} 1-\lambda & \lambda & 0 & \cdot & 0 \\ 0 & 1-\lambda & \lambda & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & 1-\lambda & \lambda \\ 0 & \cdot & \cdot & 0 & 1 \end{bmatrix},$$
on définit les matrices M^{uy} par :
$$M^{uy} = \begin{cases} M & \text{si } u = 0, y = 0\ , \\ [0_{E\times(y-1)}, M_{.y}, 0_{E\times(E-y)}] & \text{si } u = 1,\ y \in \mathcal{E}\ , \\ [\mathbf{1}, 0_{E\times E-1}] & \text{si } u = 2,\ y = 1\ , \\ 0_{E\times E} & \text{ailleurs .} \end{cases}$$
6. La loi initiale est :
$$P_x^y = \begin{cases} 1 & \text{si } x = 1,\ y = 0\ , \\ 0 & \text{sinon .} \end{cases}$$

7. Le coût instantané est défini par :

$$c_x^u = \begin{cases} \pi_x & \text{si } u = 0,\ x \in \mathcal{E}\ , \\ \tau & \text{si } u = 1,\ \forall x \in \mathcal{E}\ , \\ \rho & \text{si } u = 2,\ \forall x \in \mathcal{E}\ . \end{cases}$$

On veut minimiser le coût de maintenance sur un horizon infini $\mathbb{E}(J(z))$ dans l'ensemble des stratégies $z \in \mathcal{Z}$ où :

- les stratégies $z : \mathcal{S}_E \mapsto \mathcal{F}$ dépendent du passé des observations par l'intermédiaire du filtre optimal $Q^k = \mathbb{P}(X^k \mid Y^0, \cdots, Y^k)$ (on note $Z^k = z(Q^{k-1})$ la décision prise à l'instant k),
- le coût est défini par :

$$J(z) = \lim_{\lambda \to 0} \lambda \sum_{k=0}^{\infty} \frac{c_{X_k}^{Z_k}}{(1+\lambda)^{(k+1)}}\ .$$

Si on suppose que les coûts sont tels qu'on a intérêt à remplacer l'appareil lorsqu'il est cassé, on peut voir, que la chaîne de Markov représentant le filtre optimal a une seule classe finale. La fonction valeur $v(q) = \min \mathbb{E}(J(z) \mid Q^0 = q)$ ne dépend pas de q elle est l'unique solution notée v de l'équation de la programmation dynamique en les inconnues (v, w)

$$v + w(q) = \min \left\{ w(qm) + q.\pi, \sum_{\chi \in \mathcal{E}} w(e^\chi) q_\chi + \tau, w(e^1) + \rho \right\}\ ,$$

où

$$e_x^\chi = \begin{cases} 1 & \text{si } \chi = x,\ \ \chi, x \in \mathcal{E}, \\ 0 & \text{sinon .} \end{cases}$$

La fonction w est définie à une constante additive près.

Il est clair que le support de la loi du filtre est l'ensemble

$$\mathcal{D} = \{q = e^\chi M^\eta \mid \chi \in \mathcal{E},\ \eta \in \mathbb{N}\}\ ,$$

et donc que l'on peut paramétriser cet ensemble par les deux paramètres χ (dernier état observé), η (nombre de pas de temps passés depuis la dernière observation de l'état). L'équation de la programmation s'écrit alors en utilisant cette paramétrisation des états du filtre :

$$v + w(\chi, \eta) = \min \left\{ w(\chi, \eta + 1) + e^\chi M^\eta \pi,\ \sum_{\chi' \in \mathcal{E}} w(\chi', 0) e^\chi M^\eta_{\chi'} + \tau,\ w(1,0) + \rho \right\}\ .$$

4.2 Résolution numérique

Pour pouvoir résoudre cette équation de la programmation dynamique il suffit de borner la variable $\eta \in \mathbb{N}$. Nous le ferons en faisant l'approximation $qM^k = e^E$ pour tout q tout $k \geq N$ où N est un nombre donné à l'avance. Si l'on suppose alors que les données sur les coûts sont tels qu'il faut remplacer le matériel dès qu'il est cassé. On a le système à résoudre :

$$\begin{aligned} v^N + w^N(\chi,\eta) = \min \Big\{ & w^N(\chi,\eta+1) + e^{\chi} M^{\eta} \pi, \\ & \sum_{\chi' \in \mathcal{E}} w^N(\chi',0) e^{\chi} M^{\eta}_{\chi'} + \tau, \\ & w^N(1,0) + \rho \Big\} \ , \\ w^N(\chi,N) = w^N(1,0) + \rho \ . & \end{aligned}$$

La figure (3) donne alors les résultats numériques obtenus sur un exemple.

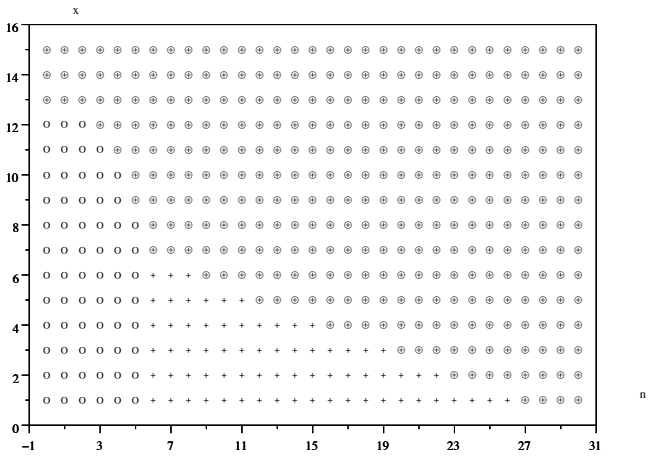

FIG. 3 – Stratégie optimale pour le probléme de maintenance dans le cas : $E = 15$, $N = 30$, $\lambda = 0.3$, $\tau = 30$, $\pi_x = x$, et $\rho = 150$. Le signe o [resp +] [resp. $\oplus$] correspond à la commande $u = 0$ [resp. $u = 1$] [resp. $u = 2$].

La figure (3) montre que si le temps passé sans observer l'état de l'appareil n'est pas très grand il vaut mieux ne rien faire (ni test ni remplacement). Si le dernier état observé n'était pas trop mauvais on a intérêt à tester l'appareil sans remplacement pendant un certain temps, puis à partir d'un temps plus long de remplacer l'appareil.

5 Conclusion

L'étude de deux exemples (notamment le premier), nous a permis de montrer que la résolution numérique des problèmes de commande optimale stochastique pour des chaînes de Markov incomplètement observées n'est pas toujours difficile. A priori la résolution de ces problèmes de commande nécessite la discrétisation d'un simplexe de dimension le nombre d'états de la chaîne de Markov. Cette tâche conduit très vite à des tailles prohibitives de la mémoire utilisée.

Le point de vue adopté ici (qui a permis de résoudre le problème de la secrétaire en dimension 200 avec 10 bits de précision) consiste à explorer numériquement le support de la loi du filtre optimal. Il se révèle particulièrement petit dans le cas du problème de la secrétaire (1 millier de points du simplexe sont suffisants pour approximer ce support dans le cas des 200 secrétaires). Ce problème se résout alors en des temps de l'ordre de la minute. La raison de ce phénomène est clair : le filtre optimal a une seule classe finale (le sommet du simplexe correspondant à une probabilité de 1 sur la meilleure secrétaire). Toutes les réalisations des trajectoires du filtre convergent vers ce point.

Le problème de maintenance est encore plus simple puisque l'on peut trouver un codage en dimension deux du support de la loi du filtre dans ce cas.

Ces exemples, relativement intéressants pratiquement, sont encourageants mais restent trop simples pour se faire une idée générale. Dans chaque cas particulier, il est nécessaire d'étudier numériquement ou théoriquement ce support avant de pouvoir avoir une idée précise de la difficulté du problème. Par contre, la complexité à priori, basée sur la taille du simplexe, n'a pas grand sens.

Références

[1] K.J. Astrom *Optimal Control of Markov Processes with Incomplete State Estimation*, Journal of Mathematical Analysis and Applications, Vol-10, pp. 174-205, 1965.

[2] Miklos Ajtai, Nimrod Megiddo et Orli Waarts : *Improved Algorithms and Analysis for Secretary Problems and Generalizations*, http ://theory.stanford.edu/ megiddo/pdf/secrfin.pdf, July 19, 1995.

[3] Robert W. Chen, Burton Rosenberg and Larry A. Shepp : *A secretary problem with two decision makers* Journal of Applied Probability 34 (1997) 1068-1074. http ://jackson.cs.miami.edu/ burt/papers/1997.0/crsf.pdf

[4] P.R. Freeman : *The secretary probleme and its Extensions.* International Statistical Review, 51, 189-206, 1983. http ://theory.stanford.edu/ megiddo/pdf/secrfin.pdf

[5] Ben Gum : *The Secretary Problem with a Uniform Distribution*, Tenth SIAM Conference on Discrete Mathematics, June 2000.

[6] M. Hauskrecht : *Value-function approximations for partially observable Markov decision processes* Journal of Artificial Intelligence Research, vol.13, pp. 33-94, 2000.

[7] Shoou-Ren Hsiau and Jiing-Ru Yang : *A natural variation of the standard secretary problem*, National Changhua University of Education, Statistica Sinica 1, 639-646, 2000.

[8] M.D. Lee, T.A. O'Connor and M.B. Welsh : *Decision making on the Full Information secretary Problem*, COGSI, Chicago, 2004. http ://www.cogsci.northwestern.edu /cogsci2004/papers/paper279.pdf

[9] V. Makis and X. Jiang *Optimal Replacement Under Partial Observations* Mathematics of Operations Research, Vol 23, No 2, pp.382-394, 2003.

[10] Alain Pagès, Michel Gondran : *Fiabilité des systèmes,* Eyrolles, 1980.

[11] Jean-Pierre Quadrat & Michel Viot : *Introduction à la commande stochastique*, http ://www-rocq.inria.fr/metalau/quadrat, 1991.

[12] M. P. Quine and J. S. Law *Exact results for a secretary problem*, Journal of Applied Probability 33, 630-639, 1996. http ://www.maths.usyd.edu.au :8000/u /malcolmq/abstracts/QL1.html

[13] SCILAB : http ://scilabsoft.inria.fr/.

Table des matières

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399