Chapitre 7: Modèles sur Horizon Infini

Fabian Bastin

DIRO, Université de Montréal

IFT-6521 – Hiver 2011

Horizon infini

Nombre illimité d'étapes, système stationnaire:
U_k, X_k, g_k, f_k, et $\mathbb{P}_k$ sont les mêmes pour tout k.

à l'étape k, on observe l'état x_k, on prend une décision $u_k \in U(x_k)$, puis une variable aléatoire w_k est générée selon la loi $\mathbb{P}(\cdot \mid x_k, u_k)$. On paye un coût d'espérance $g(x_k, u_k)$, et l'état à la prochaine étape est $x_{k+1} = f(x_k, u_k, w_k)$.

Nous avons éliminé le paramètre w_k de la fonction g.
Équivaut à remplaçer $g(x_k, u_k, w_k)$ par $\mathbb{E}[g(x_k, u_k, w_k)|x_k, u_k]$.

Modèles en temps discret

Coût espéré total sur horizon infini, avec facteur d'actualisation $\alpha \leq 1$, pour une politique $\pi = (\mu_0, \mu_1, \dots)$:

$$J_\pi(x_0) \quad = \quad \lim_{N\to\infty} \mathbb{E}\left[\sum_{k=0}^{N-1} \alpha^k g(x_k, u_k)\right]$$

où $u_k = \mu_k(x_k)$ et $x_{k+1} = f(x_k, u_k, w_k)$ pour tout k.
Si le taux d'intérêt par étape est r, alors $\alpha = 1/(1+r)$.
Le coût espéré total optimal sur horizon infini:

$$J^*(x_0) \quad = \quad \inf_\pi J_\pi(x_0).$$

Plusieurs jeux de conditions peuvent garantir que $J^*(x_0)$ existe et est fini, qu'une politique stationnaire optimale existe, et qu'on peut les calculer.

Ces modèles peuvent en fait englober tous ceux que nous avons vus auparavant.

Exemples de telles conditions:

Si la fonction g est bornée, disons $|g(x,u,w)| \leq M$ et si $0 < \alpha < 1$, alors le coût espéré total est borné par $\sum_{k=0}^{\infty} \alpha^k M = M/(1-\alpha)$.

Si g est bornée et $\alpha = 1$, mais qu'il existe au moins un état absorbant dans lequel les coûts sont nuls et que l'on atteindra à un instant aléatoire (temps d'arrêt) T_1 tel que $\mathbb{E}[T_1] < \infty$ pour toute politique, ou pour au moins une politique si les coûts sont non négatifs, alors le coût espéré total est borné par $M\mathbb{E}[T_1]$.

Souvent, T_1 est borné par une v.a. géométrique. C'est vrai par exemple s'il existe des constantes $n_0 < \infty$ et $\rho < 1$ telles qu'à partir de n'importe quel état, on a $\mathbb{P}[T_1 > n_0] \leq \rho$.

Les modèles avec un nombre fini d'étapes entrent aussi dans ce cadre: il suffit de mettre le numéro de l'étape courante dans l'état.

Certains jeux de conditions permettent des fonctions g non bornées.

Voir DPOC, vol. 2, pour plus de détails.

Coût moyen par étape, pour une politique $\pi = (\mu_0, \mu_1, \dots)$:

$$J_\pi(x_0) \quad = \quad \lim_{N\to\infty} \mathbb{E}\left[\frac{1}{N}\sum_{k=0}^{N-1} g(x_k, u_k)\right].$$

Coût moyen optimal:

$$J^*(x_0) \quad = \quad \inf_\pi J_\pi(x_0).$$

Le coût moyen peut être non nul seulement si le coût espéré total est infini. D'habitude, J_π et J^* ne dépendent pas de x_0.

Une politique π est dite stationnaire si elle est de la forme $\pi = (\mu, \mu, \mu, \dots)$. Dans ce cas, on parlera souvent, par abus de langage, de la politique μ, et on notera J_π par J_μ.

Une politique stationnaire μ est optimale si $J_\mu(x) = J^*(x)$ pour tout x, et ϵ-optimale si $J_\mu(x) \leq J^*(x) + \epsilon$ pour tout x.

Modèles à étapes discrètes (ou événements discrets)

On observe le système à des instants aléatoires
$t_0 = 0 \leq t_1 \leq t_2 \leq \ldots$ (les instants des événements).

Soit $N(t) = \sup\{k : t_k \leq t\}$ = nombre d'événements durant $(0, t]$.

à l'instant t_k, le système "saute" dans l'état x_k, on l'observe et on prend une décision u_k, on paye un coût dont l'espérance est $g(x_k, u_k)$, et la paire $w_k = (t_{k+1} - t_k, x_{k+1})$ est générée selon une loi de probabilité qui dépend de (x_k, u_k).

Coût espéré total actualisé sur horizon fini t, avec taux d'actualisation $\rho > 0$ (i.e., facteur d'actualisation $e^{-\rho t}$ pour une période de temps t), pour une politique $\pi = (\mu_0, \mu_1, \dots)$:

$$J_{\pi,t}(x_0) \quad = \quad \mathbb{E}\left[\sum_{k=0}^{N(t)} e^{-\rho t_k} g(x_k, u_k)\right]$$

où $u_k = \mu_k(x_k)$. Si pas d'actualisation: $\rho = 0$.

Dans le cas de l'horizon fini, on suppose que la valeur courante de t_k est incluse dans l'état x_k, si nécessaire.

Parfois, un coût est cumulé de façon continue, à un taux $\tilde{g}(x, u, w)$ si on est dans l'état x, on a pris la décision u, et l'élément aléatoire est w. Notre formulation couvre ce cas en prenant

$$g(x_k, u_k) \quad = \quad \mathbb{E}\left[\int_{t_k}^{t_{k+1}} e^{-\rho t} \tilde{g}(x_k, u_k, w_k) dt\right].$$

Coût espéré optimal:

$$J_t^*(x_0) \quad = \quad \inf_{\pi} J_{\pi,t}(x_0).$$

Coût espéré total actualisé sur horizon infini:

$$J_\pi(x_0) \quad = \quad \lim_{t\to\infty} J_{\pi,t}(x_0).$$

Le coût espéré total optimal:

$$J^*(x_0) \quad = \quad \inf_\pi J_\pi(x_0).$$

Coût moyen par unité de temps sur horizon infini, pour une politique $\pi = (\mu_0, \mu_1, \dots)$:

$$J_\pi(x_0) \quad = \quad \lim_{t\to\infty} \mathbb{E}\left[\frac{1}{t}\sum_{k=0}^{N(t)} g(x_k, u_k)\right].$$

Le coût moyen optimal:

$$J^*(x_0) \quad = \quad \inf_\pi J_\pi(x_0).$$

Le coût moyen peut être non nul seulement si le coût espéré total est infini. D'habitude, J_π et J^* ne dépendent pas de x_0.

Temps discret

Pour le modèle sur horizon infini, on définit la fonction de coût anticipé optimal pour un modèle à horizon tronqué: $J_0(x) = 0$ et

$$\begin{aligned} J_k(x) &= \text{coût espéré total optimal si on est dans l'état } x \\ &\quad \text{et s'il ne reste que } k \text{ étapes} \\ &= \min_{u \in U(x)} \mathbb{E}_w \left[g(x,u) + \alpha J_{k-1}(f(x,u,w)) \right] \quad \text{pour } k > 0. \end{aligned}$$

On s'attend à ce que $J_k \to J^*$ lorsque $k \to \infty$, et que

$$\mu^*(x) = \arg \min_{u \in U(x)} \mathbb{E}_w \left[g(x,u) + \alpha J^*(f(x,u,w)) \right]$$

définisse une politique optimale. On va montrer que sous certaines conditions, cela est vrai et que l'on peut borner $|J_k(x) - J^*(x)|$ et $\sup_{x \in X} |J_k(x) - J^*(x)|$, et le taux de convergence de cette erreur vers 0.

On va supposer (sauf lorsqu'on dira le contraire) que g est bornée et que w prend ses valeurs dans un ensemble D dénombrable.

Notation. Soit $\mathcal{B}(X)$ l'ensemble des fonctions bornées $V : X \to \mathbb{R}$. La norme sup d'une fonction $V \in \mathcal{B}(X)$ est définie par

$$\|V\| = \|V\|_\infty = \sup_{x \in X} |V(x)|.$$

On définit les applications $T : \mathcal{B}(X) \to \mathcal{B}(X)$ et $T_\mu : \mathcal{B}(X) \to \mathcal{B}(X)$, pour une politique stationnaire μ, par

$$\begin{aligned} T(J)(x) &= \min_{u \in U(x)} \mathbb{E}_w \left[g(x,u) + \alpha J(f(x,u,w)) \right], \\ T_\mu(J)(x) &= \mathbb{E}_w \left[g(x,\mu(x)) + \alpha J(f(x,\mu(x),w)) \right]. \end{aligned}$$

Ce sont les opérateurs de la PD. Les images de T et T_μ sont dans $\mathcal{B}(X)$ car g est bornée. On peut composer ces applications:

$$\begin{aligned} T^k(J) &= T(T^{k-1}(J)), \\ T_\mu^k(J) &= T_\mu(T_\mu^{k-1}(J)), \\ T_{\mu_i} T_{\mu_{i+1}} \cdots T_{\mu_{k-1}}(J) &= T_{\mu_i}(T_{\mu_{i+1}} \cdots T_{\mu_{k-1}})(J). \end{aligned}$$

On note $J \leq J'$ si $J(x) \leq J'(x)$ pour tout $x \in X$.

Définition. Une application $\mathcal{T} : \mathcal{B}(X) \to \mathcal{B}(X)$ est dite contractante en m étapes de module ρ si

$$\|\mathcal{T}^m(J) - \mathcal{T}^m(J')\| \leq \rho \|J - J'\|,$$

pour tout $J, J' \in \mathcal{B}(X)$, pour une constante $\rho < 1$.
Si $m = 1$, on dit simplement que $\mathcal{T}$ est contractante.

Théorème du point fixe pour les applic. contractantes.
Si $\mathcal{T} : \mathcal{B}(X) \to \mathcal{B}(X)$ est contractante en m étapes alors il existe un et un seul $J^* \in \mathcal{B}(X)$ tel que $\mathcal{T}(J^*) = J^*$, i.e., $\mathcal{T}$ possède un point fixe unique dans $\mathcal{B}(X)$.
De plus, pour tout $J \in \mathcal{B}(X)$, $\lim_{k\to\infty} \|\mathcal{T}^k(J) - J^*\| = 0$.

On montre ici que les applications T et T_μ de la programmation dynamique sont monotones et contractantes.

Proposition (monotonicité).
Si $J \leq J'$, alors $T^k(J) \leq T^k(J')$ et $T_\mu^k(J) \leq T_\mu^k(J')$, pour $k \geq 1$.

Preuve: Découle directement de la définition + induction sur k. □

Proposition: (contraction).
Si $\alpha < 1$, J et J' sont dans $\mathcal{B}(X)$, et μ est une politique stationnaire, alors, pour tout $k \geq 1$, on a

$$\|T^k(J) - T^k(J')\| \leq \alpha^k \|J - J'\|,$$
$$\|T_\mu^k(J) - T_\mu^k(J')\| \leq \alpha^k \|J - J'\|.$$

Preuve: Par la proposition précédente, comme $\|\cdot\| = \|\cdot\|_\infty$,

$$
\begin{aligned}
& J' \leq J + \|J - J'\| \\
\Rightarrow \quad & T(J') \leq T(J + \|J - J'\|) = T(J) + \alpha\|J - J'\| \\
\Rightarrow \quad & T^2(J') \leq T(T(J) + \alpha\|J - J'\|) = T^2(J) + \alpha^2\|J - J'\| \\
& \vdots \qquad\qquad \vdots \\
\Rightarrow \quad & T^k(J') \leq T(T^{k-1}(J) + \alpha^{k-1}\|J - J'\|) = T^k(J) + \alpha^k\|J - J'\| \\
\Rightarrow \quad & T^k(J') - T^k(J) \leq \alpha^k\|J - J'\|
\end{aligned}
$$

et on peut répéter le même argument en permutant J et J'. On fait la même chose avec T_μ. $\square$

Corollaire. (a) J^* et J_μ sont les solutions uniques, dans $\mathcal{B}(X)$, des équations fonctionnelles (de Bellman):

$$J^* = T(J^*) \quad \text{et} \quad J_\mu = T_\mu(J_\mu).$$

(b) Pour tout $J \in \mathcal{B}(X)$,

$$\lim_{k\to\infty} \|T^k(J) - J^*\| = \lim_{k\to\infty} \|T_\mu^k(J) - J_\mu\| = 0.$$

(c) Une politique station. μ est optimale ssi $T_\mu(J^*) = T(J^*)$.

Preuve: Les deux premiers items découlent directement du théorème du point fixe. (DPOC donne aussi une preuve directe.) Pour le dernier item, on a $T_\mu(J^*) = T(J^*) = J^*$ ssi $J^* = J_\mu$ (par le théorème du point fixe), ssi μ est optimale. $\square$

On voit qu'une politique optimale stationnaire existe ssi il existe un μ qui fait atteindre le minimum dans l'équation de Bellman.

Approximations successives (itération des valeurs)

ALGORITHME AS;

$k \leftarrow 0$;

Choisir $\epsilon > 0$ et $J_0 \in \mathcal{B}(X)$ (première approximation de J^*);

RÉPÉTER

$k \leftarrow k+1$; $J_k \leftarrow T(J_{k-1})$;

TANT QUE $\|J_k - J_{k-1}\| > \epsilon$;

RETOURNER $\tilde{\mu} = \arg\min_\mu T_\mu(J_{k-1})$ comme approx. de μ^*.

Par la propriété de contraction, on a

$$\|J_k - J^*\| \leq \alpha \|J_{k-1} - J^*\| \leq \cdots \leq \alpha^k \|J_0 - J^*\|.$$

L'erreur diminue donc de manière géométrique (ou exponentielle) en fonction de k. De plus,

$$\|J_k - J^*\| \leq \alpha \|J_{k-1} - J^*\| \leq \alpha(\|J_k - J_{k-1}\| + \|J_k - J^*\|),$$

ce qui fournit une borne sur la distance entre J_k et J^*:

$$\|J_k - J^*\| \leq \|J_k - J_{k-1}\| \alpha/(1-\alpha).$$

On peut raffiner ces bornes comme suit. Pour $J_0 \in \mathcal{B}(X)$, posons

$$\begin{aligned}
\underline{\gamma_k} &= \inf_{x \in X} [T^k(J_0)(x) - T^{k-1}(J_0)(x)]; \\
\overline{\gamma_k} &= \sup_{x \in X} [T^k(J_0)(x) - T^{k-1}(J_0)(x)]; \\
\underline{c_k} &= \underline{\gamma_k} \alpha/(1-\alpha); \\
\overline{c_k} &= \overline{\gamma_k} \alpha/(1-\alpha).
\end{aligned}$$

Proposition. On a

$$\underline{c_k} \leq J^* - T^k(J_0) \leq \overline{c_k},$$

et ces bornes ne s'élargissent jamais, ni d'un coté ni de l'autre, lorsqu'on augmente k.

Preuve. Prenons $J = J_{k-1} = T^{k-1}(J_0)$ et $\gamma = \underline{\gamma_k}$. On a

$$\begin{aligned}
& T(J) \geq J + \gamma \\
\Rightarrow \quad & T^2(J) \geq T(J) + \alpha\gamma \geq J + \gamma + \alpha\gamma \\
\Rightarrow \quad & T^3(J) \geq T(J) + \alpha\gamma + \alpha^2\gamma \geq J + \gamma + \alpha\gamma + \alpha^2\gamma \\
& \qquad \vdots \\
\Rightarrow \quad & T^{n+1}(J) \geq T(J) + (\alpha + \alpha^2 + \cdots + \alpha^n)\gamma.
\end{aligned}$$

En prenant la limite:

$$J^* = \lim_{n\to\infty} T^{n+1}(J) \geq T(J) + \gamma\alpha/(1-\alpha) = T^k(J_0) + \underline{c_k}.$$

L'autre borne se démontre de la même manière.
Pour la preuve qu'elles ne s'élargissent pas, voir DPOC. □

Cas particulier: si $J_0 \leq J^*$, alors on aura toujours $\underline{c_k} \geq 0$.

Si μ est telle que $T_\mu(J) = T(J)$ (elle fait atteindre le minimum partout), on peut appliquer la proposition précédente avec J_μ et T_μ à la place de J^* et T, ce qui donne

$$\underline{c_k} \leq J_\mu - T_\mu(J) \leq \overline{c_k}.$$

On obtient alors

$$0 \leq J_\mu - J^* \leq (J_\mu - T(J_{k-1})) - (J^* - T(J_{k-1})) \leq \overline{c_k} - \underline{c_k}.$$

Comme

$$\underline{c_k} \leq J^* - J_k \leq \overline{c_k},$$

on a

$$J_k + \underline{c_k} \leq J^* \leq J_k + \overline{c_k}.$$

à la fin de l'algorithme, comme approximation finale de J^*, on pourra prendre par exemple la médiane

$$J_k + (\overline{c_k} - \underline{c_k})/2.$$

Approximation des opérateurs.

Souvent, lorsqu'on applique l'opérateur T ou T_μ à une fonction J, on ne peut pas calculer $T(J)$ ou $T_\mu(J)$ exactement sur tout l'espace d'états, mais seulement une approximation.
C'est le cas lorsque l'espace d'états est très grand ou infini.

Soit $\tilde{J}$ une approximation de $T(J)$, telle que

$$-\delta^- \leq T(J) - \tilde{J} \leq \delta^+.$$

En appliquant la proposition précédente avec $J = T^{k-1}(J_0)$, on obtient

$$-\delta^- + \underline{c_k} \leq J^* - \tilde{J} \leq \delta^+ + \overline{c_k}.$$

Souvent, en pratique, on connait J et $\tilde{J}$, mais pas $T(J)$.
Il faudra donc estimer δ^- et δ^+. On peut le faire, par exemple, en réévaluant $T(V)$ sur une grille plus fine, ou encore aux endroits ou on pense que l'erreur peut être importante.

On a aussi les bornes suivantes:

Proposition (L'Ecuyer, 1983): Soient J et $\tilde{J}$ dans $\mathcal{B}(X)$ tels que

$$\begin{aligned} -\delta^- \leq T(J) - \tilde{J} &\leq \delta^+ \qquad \text{et} \\ T_\mu(J) - \tilde{J} &\leq \delta_0. \end{aligned}$$

Alors

$$\begin{aligned} -\varphi(J - \tilde{J}, \delta^-) \leq J^* - \tilde{J} &\leq \varphi(\tilde{J} - J, \delta^+) \\ 0 \leq J_\mu - J^* &\leq \varphi(J - \tilde{J}, \delta^+) + \varphi(\tilde{J} - J, \delta_0), \end{aligned}$$

où

$$\varphi(V, x) = x + \frac{\alpha}{1-\alpha} \max(0, V + x) \qquad \text{pour } V \in \mathcal{B}(X).$$

Notation matricielle pour X fini.

$$\begin{aligned} X &= \{1, \ldots, n\} \\ P_{ij}(u) &= \mathbb{P}[x_{k+1} = j \mid x_k = i, u_k = u]. \end{aligned}$$

Les fonctions $J \in \mathcal{B}(X)$ sont alors des vecteurs à n dimensions:

$$J = \begin{pmatrix} J(1) \\ \vdots \\ J(n) \end{pmatrix}, \qquad T(J) = \begin{pmatrix} T(J)(1) \\ \vdots \\ T(J)(n) \end{pmatrix}.$$

Pour une politique stationnaire μ donnée, on a le vecteur de coûts

$$g_\mu = \begin{pmatrix} g(1, \mu(1)) \\ \vdots \\ g(n, \mu(n)) \end{pmatrix}$$

et la matrice des probabilités de transition

$$P_\mu = \begin{pmatrix} p_{11}(\mu(1)) & \cdots & p_{1n}(\mu(1)) \\ \vdots & \vdots & \vdots \\ p_{n1}(\mu(n)) & \cdots & p_{nn}(\mu(n)) \end{pmatrix}$$

On peut alors écrire $T_\mu(J)$ sous forme matricielle:

$$T_\mu(J) = g_\mu + \alpha P_\mu J$$

et l'équation $T_\mu(J_\mu) = J_\mu$ devient un système d'équations linéaires:

$$J_\mu = g_\mu + \alpha P_\mu J_\mu, \qquad \text{i.e.,} \qquad (I - \alpha P_\mu) J_\mu = g_\mu,$$

dont la solution est

$$J_\mu = (I - \alpha P_\mu)^{-1} g_\mu.$$

Les valeurs propres de αP_μ sont toutes dans le cercle de rayon $\alpha < 1$ dans le plan complexe. Cela implique que $(I - \alpha P_\mu)$ est inversible.

Gauss-Seidel

Dans l'algorithme AS tel qu'il est formulé, on doit conserver et utiliser J_{k-1} tant qu'on n'a pas calculé $J_k(x)$ pour tous les états x. Si l'espace d'états est fini et de cardinalité n, il faut alors réserver de la mémoire pour 2 vecteurs de taille n.

Que se passe-t-il si on utilise un seul vecteur J et que l'on utilise les nouvelles valeurs de x dès qu'on les a calculées?

C'est la méthode de Gauss-Seidel.

ALGORITHME AS-GS, pour X fini;

Choisir $J \in \mathcal{B}(X)$ (première approximation de J^*);
RÉPÉTER
  POUR CHAQUE $x \in X$ FAIRE $J(x) \leftarrow T(J)(x)$;
TANT QUE "pas satisfait";
RETOURNER $\tilde{\mu} = \arg\min_\mu T_\mu(J)$ comme approx. de μ^*.

L'énoncé $J(x) \leftarrow T(J)(x)$ modifie J en un seul point et cette nouvelle valeur de $J(x)$ sera immédiatement utilisée par la suite.

La boucle "POUR CHAQUE ..." transforme une fonction $J \in \mathcal{B}(X)$ en une nouvelle fonction, disons $F(J)$. Si les états sont $\{1, 2, \ldots, n\}$ et sont traités dans l'ordre par l'algorithme, on a

$$F(J)(i) = \min_{u \in U(i)} \left[g(i,u) + \alpha \left(\sum_{j=1}^{i-1} p_{ij}(u) F(J)(j) + \sum_{j=i}^{n} p_{ij}(u) J(j) \right) \right].$$

On définit F_μ de la même façon.

Proposition. L'opérateur $F : \mathcal{B}(X) \to \mathcal{B}(X)$ est contractant de module α, tout comme T, ce qui assure la convergence. De plus, si $J \leq T(J) \leq J^*$, alors $T^k(J) \leq F^k(J) \leq J^*$, donc dans ce cas l'erreur diminue au moins aussi vite avec F qu'avec T.

Preuve. Pour J et J' dans $\mathcal{B}(X)$, on montre par induction sur i que $|F(J)(i) - F(J')(i)| \leq \alpha \|J - J'\|$. On a

$$|F(J)(1) - F(J')(1)| \leq \alpha \max_{j \in X} |J(j) - J'(j)| = \alpha \|J - J'\|.$$

Si on suppose que $|F(J)(j) - F(J')(j)| \leq \alpha \|J - J'\|$ pour tout $j < i$, alors

$$\begin{aligned} & |F(J)(i) - F(J')(i)| \\ \leq \quad & \alpha \max \left(\max_{j<i} |F(J)(j) - F(J')(j)|, \max_{j \geq i} |J(j) - J'(j)| \right) \\ \leq \quad & \alpha \|J - J'\|. \end{aligned}$$

On a donc $\|F(J) - F(J')\| \leq \alpha \|J - J'\|$, et même chose pour F_μ.

La preuve de la deuxième partie (monotonicité) se fait facilement par induction sur i pour chaque k, puis sur k. □

Lorsqu'on utilise les bornes sur l'erreur, il n'est pas clair que les bornes convergent plus vite avec AS-GS qu'avec AS standard. Le principal avantage de AS-GS est qu'il demande moins de mémoire. D'autre part, AS est plus facile à paralléliser.

Généralisation. On peut mettre à jour les valeurs de $J(i)$ en visitant les états i de manière arbitraire, possiblement aléatoire. La convergence est assurée en autant que chaque état est visité infiniment souvent lorsque le nombre d'itérations tend vers l'infini. Une version du théorème du point fixe s'applique même si les différents états visités (i.e., la mise à jour de $J(i)$ pour les différents i) se fait de manière asynchrone. Cela facilite les implantations parallèles.

On pourrait par exemple avoir:

RÉPÉTER
  CHOISIR un i au hasard dans X;
  FAIRE $J(i) \leftarrow T(J)(i)$;
TANT QUE "pas satisfait";

Itération des politiques (IP)

ALGORITHME IP;

Choisir $\epsilon > 0$;

Choisir une politique stat. μ (première approx. de μ^*);

RÉPÉTER

Trouver J tel que $J = T_\mu(J)$; (on a $J = J_\mu$)

Trouver μ tel que $T_\mu(J) = T(J)$ (nouvelle politique);

TANT QUE $\|J - T(J)\| > \epsilon$;

RETOURNER μ.

Note: lorsqu'on cherche un nouveau μ, on se restreint aux politiques admissibles raisonnables, compte tenu de la structure du problème.

Souvent, μ change très peu d'une itération à l'autre.
Dans ce cas, seulement quelques lignes du système d'équations linéaire $J = T_\mu(J)$ vont changer et il suffira de faire une "mise-à-jour" de la solution.

Proposition. à chaque itération de cet algorithme, la fonction $J = J_\mu$ ne peut augmenter en aucun point par rapport au J précédent. Elle ne peut que diminuer. De plus, dès que J ou μ ne change pas d'une itération à la suivante, la politique μ est nécessairement optimale.

Dans le cas où le nombre total de politiques stationnaires est fini (e.g., si X et U le sont), on peut remplacer "$> \epsilon$" par "> 0" et l'algorithme s'arrête toujours après un nombre fini d'itérations et retourne une politique optimale.

Preuve. Soit ν la politique a une itération donnée et μ la politique a l'itération suivante. On veut montrer que $J_\mu \leq J_\nu$.
Par définition de μ, on a

$$T_\mu(J_\nu) = T(J_\nu) \leq T_\nu(J_\nu) = J_\nu.$$

Ainsi, par la monotonicité de T_μ, $T_\mu^{k+1}(J_\nu) \leq T_\mu^k(J_\nu)$, et, en vertu des équations de Bellman,

$$J_\mu = \lim_{k\to\infty} T_\mu^k(J_\nu) \leq J_\nu.$$

Si $J_\mu = J_\nu$, alors $J_\nu = T_\mu(J_\nu) = T(J_\nu)$ et donc $J_\nu = J^*$, car c'est le seul point fixe de T. En d'autres mots, si μ n'est pas encore optimale, on doit avoir $J_\mu(x) < J_\nu(x)$ pour au moins un état $x \in X$.

Tant que l'algorithme ne s'arrête pas, il améliore nécessairement la politique à chaque itération. Donc le nombre d'itérations ne peut pas dépasser le nombre total de politiques stationnaires. □

Algorithme d'IP modifié

En pratique, lorsque X est très grand, on ne peut résoudre l'équation $J = T_\mu(J)$ qu'approximativement à chaque itération. L'une des façons de faire cela est d'utiliser l'algorithme des approximations successives pour un nombre fini d'itérations, disons m_k itérations lors du k-ième tour de boucle de l'algorithme IP. On remplace alors "Trouver J ..." par "$J \leftarrow T_\mu^{m_k}(J)$".

On peut montrer qu'en autant que les m_k sont tous positifs, la suite des fonctions J visitées par cet algorithme converge vers J^* dans le sens que $\|J - J^*\| \to 0$, et que si le nombre de politiques stationnaires est fini, alors après un nombre fini k^* d'itérations toutes les politiques visitées seront optimales. Par contre, J ne sera peut-être jamais exactement égal à J^*.

On peut aussi approximer J_μ d'une autre façon, puis choisir une prochaine politique μ telle que $T_\mu(J)$ est "proche" de $T(J)$. Lorsqu'on décide de s'arrêter, on peut calculer les mêmes bornes sur l'erreur que pour AS.

DPOC (Vol. 2, Proposition 1.3.6) nous dit ce qui se passe à la limite lorsque l'erreur d'approximation est bornée par le même constante à toutes les itérations.

Proposition. Si à chaque itération de IP, on trouve J tel que $\|J - T_\mu(J)\| \leq \delta$, puis μ tel que $\|T_\mu(J) - T(J)\| \leq \epsilon$, alors

$$\limsup_{k \to \infty} \|J_\mu - J^*\| \leq \frac{\epsilon + 2\alpha\delta}{(1-\alpha)^2}.$$

Par contre, cette proposition ne donne pas de bornes sur $J_\mu - J^*$ à une itération donnée.

Programmation linéaire

Si $J \leq J^*$, alors $J \leq T(J)$, et vice-versa.
On voit que J^* est le "plus grand" J tel que $J \leq T(J)$.
En d'autres mots, si $|X| = n < \infty$, le vecteur J^* est la solution optimale du problème de programmation linéaire:

$$\begin{aligned} \text{maximiser} \quad & J(1) + \cdots + J(n) \\ \text{s.l.c.} \quad & J(i) \leq g(i,u) + \alpha(P_{i1}(u)J(1) + \cdots + P_{in}(u)J(n)) \\ & \text{pour } i = 1, \ldots, n, \quad u \in U(i). \end{aligned}$$

Maximiser la somme revient à rendre chaque contrainte active.

Ce problème possède n variables et $U(1) + \cdots + U(n)$ contraintes. On le résoudra habituellement par une méthode duale, car il a y beaucoup moins de variables que de contraintes. Mais la résolution devient très difficile (ou impossible) si n est trop grand.

Chaque politique μ correspond à une base réalisable du PL dual, et vice-versa. La matrice de base correspondante est $I - \alpha P_\mu$ et on a $\mu(i) = u$ ssi la variable duale correspondante est strictement positive.

Les contraintes qui sont satisfaites à égalité par la solution optimale correspondent aux paires (i, u) telles que la décision u est optimale dans l'état i.

Si dans la méthode IP, on ne change la politique à chaque itération que pour l'état i pour lequel $|J(i) - T(J)(i)|$ est le plus grand, alors cette méthode est équivalente, pivot pour pivot, à appliquer l'algorithme dual du simplexe au PL.

Exemple: remplacement d'un équipement

Une machine (ou un équipement) est dans l'un des états $\{1, \ldots, n\}$. Plus l'état est élevé, plus la détérioration est avancée.Si l'état est i au début d'une période, le coût d'opération (espéré) pour cette période est $g(i)$ et l'état sera j au début de la prochaine période avec probabilité p_{ij}. Au début de chaque période, on peut laisser la machine pour une autre période ($u = 0$), ou encore la remplacer par une neuve ($u = 1$) au coût R, auquel cas la machine sera dans l'état 1 (neuve) et y restera au moins jusqu'à la prochaine période. Les coûts sont actualisés par un facteur α par période.

Soit $J^*(i)$ le coût espéré total optimal actualisé, sur horizon infini, à partir de maintenant, si on est dans l'état i. L'équation de la PD:

$$J^*(i) = \min \left[R + g(1) + \alpha J^*(1),\ g(i) + \alpha \sum_{j=1}^{n} p_{ij} J^*(j) \right],\ 1 \leq i \leq n.$$

La politique optimale: remplacer ssi $J^*(i) = R + g(1) + \alpha J^*(1)$.
On peut calculer (approx.) J^* par l'un des algorithmes: AS, IP, etc.

Hypothèse D: $g(i)$ est croissant en i et pour chaque j fixé,

$$\mathbb{P}[x_{k+1} \geq j \mid x_k = i, u = 0] = p_{ij} + \cdots + p_{in}$$

est croissant en i. $\square$

Cette hypothèse dit qu'une machine plus dégradée ne coûte pas moins cher, et a au moins autant de chances d'atteindre un seuil de dégradation donné à la prochaine étape, qu'une machine moins dégradée.

Proposition. Sous l'hypothèse D, si $J_0 = 0$ et $J_k = T^k(J_0)$, alors $J_k(i)$ est croissante en i, et $J^*(i) = \lim_{k\to\infty} J_k(i)$ est aussi croissante en i. La politique optimale est donc déterminée par un seuil i^*: on remplace ssi

$$i \geq i^* \stackrel{\text{def}}{=} \inf\{i : J^*(i) = R + g(1) + \alpha J^*(1)\}.$$

Si cet ensemble est vide, on pose $i^* = \infty$.

Preuve. On montre par induction sur k que J_k est croissante. Vrai pour $k = 0$. Supposons que c'est vrai pour $k - 1$. On a

$$J_k(i) = \min \left[R + g(1) + \alpha J_{k-1}(1),\ g(i) + \alpha \sum_{j=1}^{n} p_{ij} J_{k-1}(j) \right].$$

On peut réécrire la somme, en posant $J_{k-1}(0) = 0$, comme:

$$\sum_{j=1}^{n} p_{ij} J_{k-1}(j) = \sum_{j=1}^{n} (p_{ij} + \cdots + p_{in})(J_{k-1}(j) - J_{k-1}(j-1)).$$

L'hypothèse D nous assure que cette somme et $g(i)$ sont croissants en i, car $J_{k-1}(j) - J_{k-1}(j-1) \geq 0$ par l'hypothèse d'induction. $\square$

Coût total, $\alpha = 1$ (plus court chemin stochastique)

Si $\alpha = 1$, il faut faire des hypothèses garantissant que les coûts ne s'accumulent pas à l'infini.
On va supposer ici que $X = \{1, \ldots, n, t\}$ et que chaque $U(i)$ est fini.
L'état t est un état terminal (absorbant) dans lequel le coût est nul.
On a alors un problème de plus court chemin stochastique.

Définition. Une politique μ est propre s'il existe $n_0 < \infty$ tel que

$$\rho_\mu \stackrel{\text{def}}{=} \max_{i \in X} \mathbb{P}[x_{n_0} \neq t \mid x_0 = i, \mu] < 1.$$

Autrement, elle est impropre.

Hypothèse PP. Il existe au moins une politique stationnaire propre, et pour toute politique impropre, $J_\mu(i) = \infty$ pour au moins un i.

Sous l'hypothèse PP, T et T_μ ne sont pas nécessairement des applications contractantes par rapport à la norme sup, mais on peut quand même montrer:

Proposition. (a) Si μ est propre, alors $J = J_\mu$ est l'unique solution de l'équation $J = T_\mu(J)$, et pour tout J on a

$$J_\mu = \lim_{k \to \infty} T_\mu^k(J).$$

(b) Si $T_\mu(J) \leq J$ pour au moins un J, alors μ est propre.
(c) $J = J^*$ est l'unique solution de l'équation de Bellman $J = T(J)$, et pour tout J on a

$$J^* = \lim_{k \to \infty} T^k(J).$$

(d) Une politique stationnaire μ est optimale ssi $T_\mu(J^*) = T(J^*)$.

Preuve. Pour (a), si μ est propre, on a, avec la notation vectorielle,

$$T_\mu^k(J) = g_\mu + P_\mu T_\mu^{k-1}(J) = \cdots = P_\mu^k J + \sum_{\ell=0}^{k-1} P_\mu^\ell g_\mu.$$

Mais lorsque $k \to \infty$, $P_\mu^k J \leq \rho_\mu^{\lfloor k/n_0 \rfloor} \|J\| \to 0$ et donc

$$T_\mu^k(J) \to \sum_{\ell=0}^{\infty} P_\mu^\ell g_\mu = J_\mu.$$

On a aussi $T_\mu^{k+1}(J) = g_\mu + P_\mu T_\mu^k(J)$, qui devient, lorsque $k \to \infty$,

$$J_\mu = g_\mu + P_\mu J_\mu = T_\mu(J_\mu).$$

D'autre part, si $J = T_\mu(J)$, alors $J = \lim_{k\to\infty} T_\mu^k(J) = J_\mu$, ce qui démontre l'unicité.

(b) Si $T_\mu(J) \leq J$, puisque T_μ est monotone,

$$P_\mu^k J + \sum_{\ell=0}^{k-1} P_\mu^\ell g_\mu = T_\mu^k(J) \leq J.$$

Ceci implique que μ est propre, car autrement au moins une des composantes du vecteur défini par la somme à gauche devrait diverger vers $+\infty$ selon notre hypothèse PP.

(c) et (d): voir DPOC. $\square$

Cette proposition tient même si les $U(i)$ ne sont pas finis, par ex. si les $U(i)$ sont compacts et si les $P_{ij}(u)$ et $g(i,u)$ sont continus en u.

La proposition implique que l'algorithme AS converge vers J^*.
Il en est de même pour AS-GS.

S'il existe une politique optimale μ^* sans cycle (i.e., on ne revient jamais à un état déjà visité, ce qui implique en particulier que $P_{ii}[\mu^*(i)] = 0$ pour tout i), et si on démarre avec $J_0 = \infty$, alors l'algorithme AS atteint J^* (et ne bouge plus par la suite) après un nombre fini d'itérations.

L'algorithme IP exact converge si on se limite à explorer des politiques propres.

Pour l'algorithme IP modifié, on peut aussi adapter la preuve de convergence si on suppose que J_0 satisfait $T(J_0) \leq J_0$.

Exemple. On a 2 états: 0 et 1. L'état 0 est terminal.
Dans l'état 1, on choisit une décision $u \in (0, 1]$. Puis, avec probabilité $1 - u^2$ on fait un gain de u et on reste dans l'état 1, et avec probabilité u^2 on fait un gain de 0 et on passe à l'état terminal.
Une politique stationnaire μ correspond à choisir une valeur de $u = \mu(1)$, et une telle politique est toujours propre.
On cherche une politique qui maximise le gain espéré total.
Interprétation: on peut voir u comme la "prime de protection" demandée à une victime à chaque mois par une organisation criminelle. On passe à l'état terminal lorsque la victime refuse de céder. On a ici

$$J_\mu(1) = (1 - u^2)(u + J_\mu(1))$$

dont la solution est $J_\mu(1) = (1 - u^2)/u$.
Ainsi, $J^*(1) = \sup_{0<u\leq 1} J_\mu(1) = \infty$, mais aucune valeur de u ne permet d'atteindre cette valeur: aucune politique n'est optimale.

Si on remplaçait $U = (0, 1]$ par un nombre fini de valeurs de u positives, la proposition précédente s'appliquerait et la politique optimale serait de choisir la plus petite valeur de u.

Si on prend plutôt $U = [0, 1]$, l'hypothèse PP est violée car $u = 0$ définit une politique μ impropre pour laquelle $J_\mu(0) = J_\mu(1) = 0$.

à noter qu'il existe ici une politique non stationnaire pour laquelle le gain espéré total est infini: prenons $u_k = \mu_k(1) = 1/[4(k + 1)]$. (Faire détails.)

Fonction g non bornée.

Supposons que le nombre d'états est infini, que g n'est pas nécessairement bornée et/ou que α peut valoir 1.

Dans ce cas, les choses se compliquent car il pourrait s'accumuler, par exemple, un coût infini ou un revenu infini, ou même les deux en même temps ce qui pourrait faire un coût net indéterminé.

L'algorithme de la PD fonctionne quand même et plusieurs propriétés fonctionnent toujours sous l'une des conditions suivantes:
Hypothèse P: $g(x, u) \geq 0$ pour tout x et $u \in U(x)$.
Hypothèse N: $g(x, u) \leq 0$ pour tout x et $u \in U(x)$.

Proposition. (a) Sous P ou N, $J^* = T(J^*)$ et $J_\mu = T_\mu(J_\mu)$. Ces équations peuvent avoir d'autres solutions que J^* et J_μ.
Mais sous P, aucune solution n'est plus petite, et sous N, aucune solution n'est plus grande.
(b) Sous P, μ est optimale ssi $T(J^*) = T_\mu(J^*)$.
(c) Sous P, si $J_\infty \stackrel{\text{def}}{=} \lim_{k\to\infty} T^k(0)$ satisfait $J_\infty = T(J_\infty)$ et si $J \in \mathcal{B}(X)$ et $\alpha < 1$, ou si $0 \leq J \leq J^*$, alors $\lim_{k\to\infty} T^k(J) = J^*$.
(d) Sous P, s'il existe $\bar{k}$ tel que pour $k \geq \bar{k}$, $x \in X$ et $\lambda \in \mathbb{R}$,

$$\left\{u \in U(x) \mid g(x,u) + \alpha\mathbb{E}[T^k(J_0)(f(x,u,w))] \leq \lambda\right\}$$

est un ensemble compact, alors $J_\infty = J^*$ et il existe une politique stationnaire optimale.
(e) Sous N, μ est optimale ssi $T(J_\mu) = T_\mu(J_\mu)$.
(f) Sous N, si $J \in \mathcal{B}(X)$ et $\alpha < 1$, ou si $J^* \leq J \leq 0$, alors $\lim_{k\to\infty} T^k(J) = J^*$.

Coût moyen par étape

Considérons un modèle actualisé avec $X = \{1, \dots, n\}$ et $\alpha < 1$. Notons $J_{\alpha,\mu}$ et J^*_α les valeurs de J_μ et J^* pour un α donné. On va faire tendre α vers 1.

Coût moyen par étape, pour une politique stationnaire μ:

$$\begin{aligned}
& \lim_{N\to\infty} \mathbb{E}\left[\frac{1}{N}\sum_{k=0}^{N-1} g(x_k, \mu(x_k))\right] \\
= & \lim_{N\to\infty} \lim_{\alpha\to 1^-} \frac{\mathbb{E}\left[\sum_{k=0}^{N-1} \alpha^k g(x_k, \mu(x_k))\right]}{\sum_{k=0}^{N-1} \alpha^k} \\
= & \lim_{\alpha\to 1^-} \lim_{N\to\infty} \frac{\mathbb{E}\left[\sum_{k=0}^{N-1} \alpha^k g(x_k, \mu(x_k))\right]}{\sum_{k=0}^{N-1} \alpha^k} \\
= & \lim_{\alpha\to 1^-} (1-\alpha) J_{\alpha,\mu}(x_0)
\end{aligned}$$

si on suppose que l'on peut échanger les deux limites.

On choisit un état, disons t, comme état de référence, et on pose:

$$\begin{aligned} h_{\alpha,\mu}(i) &= J_{\alpha,\mu}(i) - J_{\alpha,\mu}(t), \\ \lambda_{\alpha,\mu} &= (1-\alpha) J_{\alpha,\mu}(t), \\ h_\mu(i) &= \lim_{\alpha \to 1^-} h_{\alpha,\mu}(i) \quad \text{et} \\ \lambda_\mu &= \lim_{\alpha \to 1^-} \lambda_{\alpha,\mu} \end{aligned}$$

si ces deux limites existent. Les valeurs $h_{\alpha,\mu}(i)$ et $h_\mu(i)$ représentent un coût différentiel de l'état i par rapport à l'état de référence, tandis que λ_μ représente le coût moyen par étape sur horizon infini, sous la politique μ, pour $x_0 = t$. En fait, sous l'hypothèse que tous les états communiquent, λ_μ ne dépend pas de x_0.

L'équation $J_{\alpha,\mu} = g_\mu + \alpha P_\mu J_{\alpha,\mu}$ se réecrit:

$$\begin{aligned} h_{\alpha,\mu} + J_{\alpha,\mu}(t) &= g_\mu + \alpha P_\mu (h_{\alpha,\mu} + J_{\alpha,\mu}(t)) \\ \lambda_{\alpha,\mu} + h_{\alpha,\mu} &= g_\mu + \alpha P_\mu h_{\alpha,\mu}, \end{aligned}$$

qui devient, lorsque $\alpha \to 1^-$,

$$\lambda_\mu + h_\mu = g_\mu + P_\mu h_\mu \stackrel{\text{def}}{=} T_\mu(h_\mu). \quad (1)$$

De la même manière, pour les valeurs optimales, on pose

$$\begin{aligned} h^*_\alpha(i) &= J^*_\alpha(i) - J^*_\alpha(t), \\ \lambda_\alpha &= (1-\alpha) J^*_\alpha(t), \\ h^*(i) &= \lim_{\alpha \to 1^-} h^*_\alpha(i) \quad \text{et} \\ \lambda^* &= \lim_{\alpha \to 1^-} \lambda_\alpha \end{aligned}$$

si ces deux limites existent. Les fonctions h^*_α et h^* représentent le coût différentiel à l'optimum, et λ^* le coût moyen optimal par étape, sur horizon infini. Dans la plupart des cas, ce coût moyen ne dépend pas de l'état initial x_0.

L'équation de Bellman $J^*_\alpha = \min_\mu \{g_\mu + \alpha P_\mu J^*_\alpha\}$ se réécrit

$$\lambda_\alpha + h^*_\alpha = \min_\mu \{g_\mu + \alpha P_\mu h^*_\alpha\}$$

et devient à la limite

$$\lambda^* + h^* = \min_\mu (g_\mu + P_\mu h^*) \stackrel{\text{def}}{=} T(h^*). \qquad (2)$$

Le raisonnement que nous venons de faire est heuristique, à cause des hypothèses que nous avons faites sur les limites.
S'il tient, (1) et (2) donnent des conditions nécessaires d'optimalité.
Sont-elles aussi suffisantes? Et si les états ne communiquent pas tous sous certaines politiques?
Les propositions qui suivent répondent à ces questions.

Proposition.
Il existe une constante λ et une fonction $h \in \mathcal{B}(X)$ telles que

$$\lambda + h = T(h)$$

si et seulement si $\lambda = J^*(i) = \min_\pi J_\pi(i)$ (le coût moyen optimal) pour tout $i \in X$.
Et si $T_\mu(h) = T(h)$ pour ce h, alors μ est optimale.
De même, il existe λ_μ et $h_\mu \in \mathcal{B}(X)$ tels que

$$\lambda_\mu + h_\mu = T_\mu(h_\mu)$$

si et seulement si $\lambda_\mu = J_\mu(i)$ pour tout $i \in X$.

Preuve partielle. Supposons que $\lambda + h = T(h)$.
Si $\pi = (\mu_0, \mu_1, \dots)$ est une politique admissible et $N > 0$, alors

$$\begin{aligned} T_{\mu_{N-1}}(h) &\geq T(h) = \lambda + h, \\ T_{\mu_{N-2}}(T_{\mu_{N-1}}(h)) &\geq T_{\mu_{N-2}}(\lambda + h) = \lambda + T_{\mu_{N-2}}(h) \geq 2\lambda + h, \\ &\vdots \\ T_{\mu_0} \cdots T_{\mu_{N-1}}(h) &\geq N\lambda + h \end{aligned}$$

et on a l'égalité partout si chaque μ_k fait atteindre le minimum. On a donc, sous la politique π et pour $x_0 = i$,

$$\begin{aligned} \frac{1}{N}\mathbb{E}\left[h(x_N) + \sum_{k=0}^{N-1} g(x_k, \mu(x_k))\right] &= \frac{1}{N} T_{\mu_0} \cdots T_{\mu_{N-1}}(h)(i) \\ &\geq \lambda + \frac{h(i)}{N}. \end{aligned}$$

Lorsque $N \to \infty$, cela donne $J_\pi(i) \geq \lambda$, avec l'égalité si chaque μ_k fait atteindre le minimum.

Cette preuve fonctionne même si X et Y sont infinis. La preuve dans l'autre direction dépend du fait que $|X|$ est fini (voir DPOC).

Etant donné une chaîne de Markov à états finis, avec une matrice de transition de probabilité P, une classe récurrente est un ensemble d'états qui commiquent dans le sens que de chaque état de l'ensemble, il existe une probabilité de 1 de visiter finalement tous les autres états de l'ensemble, et une probabilité nulle d'aller à un moment donné vers un état hors de l'ensemble.

Une politique μ est dite unichaîne si elle donne lieu à une seule classe d'états récurrents (et éventuellement certains états transitoires).

Proposition. Si μ est une politique unichaîne, alors le système d'équations

$$\lambda + h = T_\mu(h), \qquad h(n) = 0,$$

possède l'unique solution $(\lambda, h) = (\lambda_\mu, h_\mu)$, où $\lambda_\mu = J_\mu(i)$ pour tout i.

Proposition.

Supposons que l'une des trois conditions suivantes est vérifiée.

(C1) Toute politique optimale parmi les politiques stationnaires est unichaîne.

(C2) Tous les états sont accessibles les uns des autres, i.e., pour tous i, j dans X, il existe une politique stationnaire μ et $k > 0$ tels que $\mathbb{P}[x_k = j \mid x_0 = i, \mu] > 0$.

(C3) Il existe un état i_0 et des constantes $L > 0$ et $\bar{\alpha} \in (0, 1)$ tels que

$$\sup_{i \in X,\, \bar{\alpha} < \alpha < 1} |J_\alpha(i) - J_\alpha(i_0)| \leq L.$$

Alors $J^*(i)$ ne dépend pas de i et on a

$$J^*(i) = \lambda^* = \lim_{\alpha \to 1^-} (1 - \alpha) J_\alpha^*(i)$$

pour tout i, et $\lambda^* + h^* = T(h^*)$, où h^* est défini tel que précédemment, peu importe le choix de l'état t.

Exemple: remplacement d'un équipement (suite).

Même exemple, avec l'hypothèse D, mais on veut maintenant minimiser le coût moyen par période, sur horizon infini.

Ici, les politiques ne sont pas toutes unichaînes. Par exemple, si $p_{1n} = 0$, la politique stupide qui consiste à toujours remplacer sauf si on est dans l'état n donne lieu à deux classes d'états qui ne communiquent pas entre elles: $\{1, \ldots, n-1\}$ et $\{n\}$.
C2 n'est vérifiée que sous des hypothèses supplémentaires.
Mais on peut vérifier la condition C3: Pour $\alpha < 1$, on a

$$\begin{aligned} J_\alpha^*(i) &= \min\left[R + g(1) + \alpha J_\alpha^*(1),\ g(i) + \alpha \sum_{j=1}^{n} p_{ij} J_\alpha^*(j)\right], \\ \text{d'où} \quad 0 &\le J_\alpha^*(i) - J_\alpha^*(1) \le R + g(1) + \alpha J_\alpha^*(1) - J_\alpha^*(1) \\ &\le \max\left(0,\ R - \alpha \sum_{j=1}^{n} p_{1j}(J_\alpha^*(j) - J_\alpha^*(1))\right] \le R. \end{aligned}$$

Il s'ensuit que l'équation d'optimalité

$$\lambda + h(i) = \min \left[R + g(1) + h(1),\ g(i) + \sum_{j=1}^{n} p_{ij} h(j) \right]$$

possède une solution (λ, h) et la politique optimale consiste à prendre la décision qui minimise cette expression, pour chaque i.

On a aussi que $h(i) = \lim_{\alpha \to 1^-} (J^*_\alpha(i) - J^*_\alpha(1))$ est croissant en i (en raison de l'hypothèse D), ce qui implique que la politique optimale consiste à remplacer ssi

$$i \geq i^* \stackrel{\text{def}}{=} \inf\{i : \lambda + h(i) = R + g(1) + h(1)\}.$$

Si cet ensemble est vide, on pose $i^* = \infty$.

Algorithmes de calcul

Les algorithmes pour le cas actualisé se transposent au cas du coût moyen. On suppose ici que X et U sont finis.

La méthode des approximations successives pour le vecteur des valeurs relatives h, i.e., appliquée aux équations

$$\lambda + h = T(h); \qquad h(t) = 0,$$

devient:

ALGORITHME ASR;

$k \leftarrow 0$; Choisir $\epsilon > 0$ et $t \in X$;

Choisir $h_0 \in \mathcal{B}(X)$ tel que $h_0(t) = 0$ (première approx. de h^*);

RÉPÉTER

$k \leftarrow k + 1$; $\lambda_k \leftarrow T(h_{k-1})(t)$; $h_k \leftarrow T(h_{k-1}) - \lambda_k$;

TANT QUE $\|h_k - h_{k-1}\| > \epsilon$;

RETOURNER $\tilde{\mu} = \arg\min_\mu T_\mu(h_{k-1})$ comme approx. de μ^*.

Cet algorithme ne converge pas toujours. Il peut cycler, en particulier si la suite des états visités est périodique.

La proposition suivante donne des conditions suffisantes assurant la convergence. Si ces conditions ne sont pas vérifiées, ou si on n'en est pas certain, on peut utiliser une version modifiée de l'algorithme qui consiste à remplacer la matrice P_μ par

$$\tilde{P}_\mu = \tau P_\mu + (1 - \tau) I$$

pour chaque politique stationnaire μ, où $0 < \tau < 1$.

Proposition. Supposons qu'il existe $m > 0$ tel que pour toute politique $\pi = (\mu_0, \mu_1, \dots)$, il existe $\epsilon > 0$ et un état $s \in X$ tels que tous les éléments de la colonne s des matrices $P_{\mu_m} \cdots P_{\mu_1}$ et $P_{\mu_{m-1}} \cdots P_{\mu_0}$ sont $\geq \epsilon$. Alors:

(a) La suite des h_k dans l'algorithme ASR converge vers une solution h de l'équation de Bellman $\lambda + h = T(h)$, et λ_k converge donc vers λ^*, le coût moyen optimal.

(b) Si on définit

$$\begin{aligned} \underline{c_k} &= \min_{x \in X} [T(h_k)(x) - h_k(x)]; \\ \overline{c_k} &= \max_{x \in X} [T(h_k)(x) - h_k(x)]; \end{aligned}$$

alors

$$\underline{c_k} \leq \lambda^* \leq \overline{c_k},$$

et ces bornes ne s'élargissent jamais, ni d'un coté ni de l'autre, lorsqu'on augmente k.

La version Gauss-Seidel de cet algorithme ne converge pas toujours.

Si on remplace P_μ par $\tilde{P}_\mu = \tau P_\mu + (1-\tau)I$ pour chaque μ, où $0 < \tau < 1$, l'opérateur T devient T_τ, défini par

$$\begin{aligned} T_\tau(h)(i) &= \min_{u \in U(i)} \left[g(i,u) + (1-\tau)h(i) + \tau \sum_{j=1}^{n} p_{ij}(u) h(j) \right] \\ &= (1-\tau)h(i) + \min_{u \in U(i)} \left[g(i,u) + \tau \sum_{j=1}^{n} p_{ij}(u) h(j) \right], \end{aligned}$$

i.e., $T_\tau(h) = (1-\tau)h + \min_\mu \left[g_\mu + \tau P_\mu h \right]$, et on obtient:

ALGORITHME ASR-τ;

$k \leftarrow 0$; Choisir $\epsilon > 0$, $t \in X$ et $\tau > 0$;
Choisir $h_0 \in \mathcal{B}(X)$ tel que $h_0(t) = 0$ (première approx. de h^*);
RÉPÉTER
$k \leftarrow k+1$; $\lambda_k \leftarrow T_\tau(h_{k-1})(t)$; $h_k \leftarrow T_\tau(h_{k-1}) - \lambda_k$;
TANT QUE $\|h_k - h_{k-1}\| > \epsilon$;
RETOURNER $\tilde{\mu} = \arg\min_\mu T_\mu(h_{k-1})$ comme approx. de μ^*.

Proposition. Supposons que chaque politique μ est unichaîne et que $0 < \tau < 1$. On considère la suite des constantes λ_k et des vecteurs h_k produits par l'algorithme ASR-τ.

Alors la suite des vecteurs $(\lambda_k, \tau h_k)$ converge vers un vecteur (λ, h) qui est solution de l'équation de Bellman $\lambda + h = T(h)$. On a donc $\lambda = \lambda^*$, le coût moyen optimal.

Itération des politiques (IP)

Proposition. Si on se restreint à ne considérer que les politiques μ pour lesquelles la chaine de Markov est irréductible (i.e. est unichaîne et n'a pas d'état transitoire), ou encore si on s'assure de ne jamais modifier la politique μ pour un état i tel que la décision $\mu(i)$ pour la politique précédente fait encore atteindre le minimum, alors l'algorithme IP converge en temps fini et retourne une politique optimale.

ALGORITHME IP;

Choisir $t \in X$ et $\epsilon > 0$;
Choisir une politique stationnaire μ (première approx. de μ^*);
RÉPÉTER
    Trouver (λ, h) tels que $\lambda + h = T_\mu(h)$ et $h(t) = 0$;
    (on a $h = h_\mu$)
    Trouver μ tel que $T_\mu(h) = T(h)$ (nouvelle politique);
TANT QUE $\|\lambda + h - T(h)\| < \epsilon$;
RETOURNER μ.

Modèle Semi-Markovien

Processus de renouvellement Markovien commandé (PRMC).
Le temps écoulé entre deux transitions successive est maintenant aléatoire. Soient $0 = t_0 \leq t_1 \leq t_2 \leq \cdots$ les instants des transitions (ou étapes, ou événements).
$N(t) = \sup\{k : t_k \leq t\}$ = nombre d'événements durant $(0, t]$.
à l'instant t_k, le système "saute" dans l'état x_k, on l'observe et on prend une décision u_k, et on paye un coût d'espérance $g(x_k, u_k)$.
Puis le couple $(t_{k+1} - t_k, x_{k+1})$ est généré selon la loi de probabilité conjointe $Q(\cdot \mid x_k, u_k)$.

Coût total actualisé. Les coûts sont actualisés au taux $\rho > 0$. Le coût $g(x_k, u_k)$ peut représenter en fait l'espérance du coût total cumulé sur la période $[t_k, t_{k+1})$, actualisé au temps t_k. Par exemple, si le coût est cumulé continûment au taux $c(x, u)$ quand on est dans l'état x et qu'on a pris la décision u, on aura

$$g(x_k, u_k) = \mathbb{E}\left[\int_0^{t_{k+1}-t_k} e^{-\rho\zeta} c(x_k, u_k) d\zeta\right].$$

On cherche une politique stationnaire μ qui minimise le coût espéré total actualisé sur horizon infini, pour un état initial x_0 fixé:

$$J_\mu(x_0) \quad = \quad \lim_{n\to\infty} \mathbb{E}\left[\sum_{k=0}^{n-1} e^{-\rho t_k} g(x_k, \mu(x_k)) \mid \mu, x_0\right].$$

Pour $k \geq 1$, le facteur d'actualisation espéré pour les k prochaines étapes, si on est dans l'état x et on utilise la politique μ, est

$$\alpha_k(x, \mu) = \mathbb{E}\left[e^{-\rho t_k} | x_0 = x, \mu\right].$$

En particulier, pour $k = 1$, on a

$$\alpha_1(x, \mu) = \int_0^\infty e^{-\rho\zeta} Q(d\zeta, X \mid x, \mu(x)).$$

Hypothèse C: Contraction en m étapes.
La fonction de coût g est bornée, et il existe un entier $m > 0$ et un nombre réel $\alpha < 1$ tels que

$$\sup_{x \in X,\, \mu} \alpha_m(x, \mu) \leq \alpha.$$

On définit les opérateurs de la PD:

$$\begin{aligned} T(J)(x) &= \min_{u \in U(x)} \left[g(x,u) + \mathbb{E}\left[e^{-\rho t_1} J(x_1) \mid x_0 = x, u_0 = u \right] \right] \\ &= \min_{u \in U(x)} \left[g(x,u) + \int_{[0,\infty)\times X} e^{-\rho \zeta} J(y) Q(d\zeta, dy \mid x, u) \right] \\ T_\mu(J)(x) &= g(x,\mu(x)) + \mathbb{E}\left[e^{-\rho t_1} J(x_1) \mid x_0 = x, u_0 = \mu(x) \right]. \end{aligned}$$

Sous l'hypothèse C, les opérateurs T^m et T_μ^m sont contractants de module α. On peut alors appliquer les algorithmes AS, ASG, IP, ..., comme auparavant.

Proposition. Sous l'hypothèse C, on a $T(J) = J$ ssi $J = J^*$, et $\lim_{k\to\infty} \| T^k(J) - J\| = 0$.
De même, $T_\mu(J) = J$ ssi $J = J_\mu$, et $\lim_{k\to\infty} \| T_\mu^k(J) - J\| = 0$.

De plus, les bornes sur J^* et J_μ dérivées dans le contexte du temps discret sont encore valides ici.

Exemple: Vente d'un actif. Supposons que les offres arrivent selon un processus de Poisson de taux λ (les durées entre les offres successives sont des v.a. i.i.d. exponentielles de moyenne $1/\lambda$). Les montants des offres sont des v.a. i.i.d., indép. du proc. de Poisson. Les revenus sont actualisés au taux $\rho > 0$. On veut maximiser le revenu espéré total actualisé, sur horizon infini.

Soit Δ l'état dans lequel on a vendu. Autrement, l'état x est le montant de l'offre courante. Si Z est la durée entre la date de l'offre courante et celle de la prochaine offre, alors Z est une v.a. continue de densité $\lambda e^{-\lambda\zeta}$ sur $[0, \infty)$, et les équations de récurrence s'écrivent:

$$J(x) = \begin{cases} 0 & \text{si } x = \Delta; \\ \max\{x,\ a\} & \text{sinon,} \end{cases}$$

où, si on note par w le montant de la prochaine offre,

$$a = \mathbb{E}[e^{-\rho Z} J(w)] = \int_0^\infty \lambda e^{-\lambda\zeta} e^{-\rho\zeta} \mathbb{E}[J(w)] d\zeta = \frac{\lambda}{\lambda + \rho} \mathbb{E}[\max(w, a)].$$

La politique optimale est d'accepter l'offre ssi $x > a$.

Coût moyen par unité de temps

On cherche une politique stationnaire μ qui minimise le coût moyen par unité de temps sur horizon infini.

$$J_\mu(x_0) \quad = \quad \limsup_{N\to\infty} \frac{\mathbb{E}\left[\sum_{k=0}^{N-1} g(x_k, \mu(x_k)) \mid \mu, x_0\right]}{\mathbb{E}\left[t_N \mid \mu, x_0\right]}.$$

La durée espérée jusqu'à la prochaine transition, si on est dans l'état x et on prend la décision u, est

$$\bar{\tau}(x,u) = \mathbb{E}\left[t_1 - t_0 | x_0 = x, u_0 = u\right] = \int_0^\infty \zeta Q(d\zeta, X \mid x, u).$$

Hypothèse U. Condition d'uniformisation.
La fonction g est bornée, et il existe un nombre réel $\delta > 0$ tel que

$$\inf_{x \in X,\, u \in U(x)} \bar{\tau}(x,u) > \delta.$$

Sous l'hypothèse U, on peut transformer le modèle en un modèle en temps discret équivalent dont les transitions se produisent à tous les δ unités de temps. Cela s'appelle l'uniformisation du processus. On remplace g, $\bar{\tau}$, et Q par

$$\begin{aligned}
\tilde{g}(x,u) &= g(x,u)\delta/\bar{\tau}(x,u), \\
\tilde{\tau}(x,u) &= \delta, \\
\tilde{Q}(\bullet \mid x,u) &= \frac{\delta}{\bar{\tau}(x,u)} Q([0,\infty) \times \bullet \mid x,u) + \left(1 - \frac{\delta}{\bar{\tau}(x,u)}\right) \mathbb{I}[x \in \bullet].
\end{aligned}$$

Si X est fini, si on pose $p_{ij}(u) = Q([0,\infty) \times \{j\} \mid i,u)$, et la loi de probabilité $\tilde{Q}(\cdot \mid i,u)$ correspond aux probabilités:

$$\tilde{p}_{ij}(u) = \begin{cases} p_{ij}(u)\delta/\bar{\tau}(i,u) & \text{si } j \neq i, \\ 1-(1-p_{ii}(u))\delta/\bar{\tau}(i,u) & \text{si } j = i. \end{cases}$$

On a une transition à toutes les δ unités de temps, mais elle ne change l'état qu'avec probabilité $\delta/\bar{\tau}(x,u)$. Les transitions qui laissent le système dans le même état sont des pseudo-transitions, qui ne servent qu'à uniformiser les durées entre les transitions de manière à obtenir un modèle en temps discret uniformisé.

On peut montrer que cette transformation ne change pas la valeur de $J_\mu(x_0)$: Si $\tilde{\mathbb{E}}$ représente l'espérance associée à $\tilde{Q}$, on a

$$J_\mu(x_0) \quad = \quad \limsup_{N\to\infty} \frac{1}{N\delta} \tilde{\mathbb{E}} \left[\sum_{k=0}^{N-1} \tilde{g}(x_k, \mu(x_k)) \mid \mu, x_0 \right].$$

On peut alors résoudre le modèle uniformisé par les mêmes techniques que pour le modèle en temps discret (AS, IP, ...), en utilisant les fonctions de valeurs relatives.

Les opérateurs de la PD s'écrivent:

$$\begin{aligned} T_\mu(h)(x) &= \tilde{g}(x,\mu(x)) + \int_X h(y)\tilde{Q}(dy \mid x,\mu(x)), \\ &= \frac{\delta}{\bar{\tau}(x,\mu(x))}\left(g(x,\mu(x)) + \int_X h(y)Q([0,\infty)\times dy \mid x,\mu(x))\right), \\ &\quad + \left(1 - \frac{\delta}{\bar{\tau}(x,\mu(x))}\right) h(x) \\ T(h) &= \min_\mu T_\mu(h). \end{aligned}$$

Proposition. Supposons que l'algorithme ASR converge (au sens de la norme sup) vers une solution $\tilde{h}$ de l'équation $T(h) = h + \tilde{\lambda}$ (avec $\tilde{\lambda} = T(\tilde{h})(t)$). Alors toute politique μ telle que $T_\mu(\tilde{h}) = T(\tilde{h})$ est optimale et le coût moyen optimal par unité de temps est $\lambda = \tilde{\lambda}/\delta$.

Le système d'équations $T_\mu(h) = h + \tilde{\lambda}$ se réécrit:

$$\begin{aligned} h(x) + \tilde{\lambda} &= \left(1 - \frac{\delta}{\bar{\tau}(x, \mu(x))}\right) h(x) + \frac{\delta}{\bar{\tau}(x, \mu(x))} \left(g(x, \mu(x)) \right. \\ &\qquad \left. + \int_X h(y) Q([0, \infty) \times dy \mid x, \mu(x)) \right), \end{aligned}$$

$$\begin{aligned} \tilde{\lambda} \bar{\tau}(x, \mu(x))/\delta &= -h(x) + g(x, \mu(x)) \\ &\qquad + \int_X h(y) Q([0, \infty) \times dy \mid x, \mu(x)), \end{aligned}$$

$$\begin{aligned} h(x) &= g(x, \mu(x)) - \lambda \bar{\tau}(x, \mu(x)) \\ &\qquad + \int_X h(y) Q([0, \infty) \times dy \mid x, \mu(x)). \end{aligned}$$

De la même manière, $T(h) = h + \tilde{\lambda}$ s'écrit:

$$\begin{aligned} h(x) &= \min_\mu \left[g(x, \mu(x)) - \lambda \bar{\tau}(x, \mu(x)) \right. \\ &\qquad \left. + \int_X h(y) Q([0, \infty) \times dy \mid x, \mu(x)) \right]. \end{aligned}$$

On peut utiliser cette formulation pour appliquer l'algorithme AS.

Exemple. ("The streetwalker dilemma")
On offre un certain type de service à des clients, qui arrivent selon un processus de Poisson de taux r. Pour chaque client, avec probabilité p_i, pour $i = 1, \ldots, n$, le client offre m_i dollars pour utiliser le service pendant T_i unités de temps. On a bien sûr $p_1 + \cdots + p_n = 1$. On peut rejeter l'offre ($u = 0$) ou l'accepter ($u = 1$). Toutes les offres qui arrivent lorsqu'un client utilise le service sont perdues. On veut maximiser le revenu moyen par unité de temps sur horizon infini.

Notons i l'état dans lequel on vient de recevoir l'offre (m_i, T_i). On a

$$\begin{aligned} \bar{\tau}(i,1) &= T_i + 1/r, & \text{si } g(i,1) &= m_i, \\ \bar{\tau}(i,0) &= 1/r, & \text{si } g(i,0) &= 0. \end{aligned}$$

L'équation d'optimalité devient:

$$h(i) = \max\left\{ m_i - (T_i + 1/r)\lambda + \sum_{j=1}^{n} p_j h(j),\ -\lambda/r + \sum_{j=1}^{n} p_j h(j) \right\}.$$

Politique optimale: accepter les offres qui satisfont $m_i/T_i \geq \lambda$ et rejeter les autres, où λ est le revenu optimal par unité de temps.

Exemple: Stratégie optimal d'investissement dans un contexte de crédits d'impôts

[L'Ecuyer, Haurie, Hollander 1985]
Vers 1980, aux USA et au Canada, on donnait des crédits d'impôt pour l'accroissement des dépenses en recherche et développement (R&D) par rapport à la moyenne des trois dernières années.
Question: comment une entreprise peut-elle optimiser ses dépenses de R&D dans un tel contexte?

Prenons un modèle simplifié. Supposons qu'un investissement de u dollars pour une année donnée rapporte un profit net actualisé (au début de la période) de $r(u)$. Sans les crédits d'impôt, on choisira bien sûr u qui maximise $r(u)$.

Supposons maintenant que l'entreprise reçoit un gain net additionnel de $h(x, u) = \gamma \max[0, u - (y_1 + y_2 + y_3)/3]$ où $x = (y_1, y_2, y_3)$ est le vecteur des montants investis au cours des 3 dernières années. Le prochain état sera $f(x, u) = (u, y_1, y_2)$. Le revenu net pour cette étape sera $g(x, u) = h(x, u) + r(u)$.

On veut maximiser le revenu net total actualisé, sur horizon infini: $\sum_{k=0}^{\infty} \alpha^k g(x_k, u_k)$. Il s'agit d'un problème déterministe.

Supposons que l'on impose $u \in [0, b]$. Les équations d'optimalité de la PD s'écrivent alors:

$$J(x) = \max_{0 \leq u \leq b} \left[h(x, u) + r(u) + \alpha J(f(x, u)) \right].$$

On peut résoudre cela par approx. successives ou par itération des politiques, en approximant la fonction J.

Dans l'article cité, on partitionne l'espace d'états en boites rectangulaires, et J est approximé par une fonction trilinéaire sur chaque boite rectangulaire. à chaque étape de l'algorithme AS, la fonction J est évaluée à chaque coin des boites, puis on interpole. On raffine l'approximation périodiquement, en augmentant le nombre de boites, au fur et à mesure des itérations. à la fin, on calcule les bornes sur l'erreur en estimant c^-, c^+, etc.

Exemple numérique:

$\gamma = 0.5$, $\alpha = 0.9$, $b = 4$, $r(u) = 2\ln(1+u) - u$.

Supposons que l'état initial est $x = (1, 1, 1)$.

La solution optimale (montants annuels investis):

		1.0	1.0	1.0
2.076	3.000	0.784	0.593	0.558
2.078	3.000	0.784	0.593	0.558
2.078	...			

Autres variantes: On pourrait considerer le revenu moyen par année, sur horizon infini.

Pour la résolution, on pourrait considérer un algorithme d'itération des politiques en approximant J_μ à chaque itération par une combinaison linéaire de fonctions de base.