Kolmogorov, Kuhn, Isaacs et Bellman

Pierre Bernhard
ESSI

15 août 2003

Abstract

On donne un formalisme de programmation dynamique en temps discret qui recouvre l'équation de Kolmogorov, l'induction arrière de Kuhn dans un jeu en forme extensive, l'équation d'Isaacs d'un jeu multi-étage et l'équation de Bellman de la commande stochastique.

0 Coment utiliser cette note

La plupart des utilisateurs n'auront cure d'un jeu dynamique. Par contre, en fonction de ce qu'on cherche, on peut ré-écrire tout ce qui suit de façon à le spécialiser à des cas intéressants en eux-mêmes. À savoir...

- en retirant complètement les décideurs (ce qui revient au même que de dire qu'aucun jamais n'est actif : les U^i ci-dessous sont tous réduits à des singletons), on a la forme discrète de l'équation de Kolmogorov, permettant de calculer l'espérance d'une fonctionnelle de trajectoire d'un processus markovien avec temps d'arrêt, en l'occurence celui engendré par (2) ci-dessous, dont on a retiré les u_t, avec le temps de sortie défini par (1),
- en ne laissant qu'un seul joueur, on a l'équation de Bellman d'un problème de commande stochastique en temps discret. On peut bien sûr retirer les aléas pour en faire un problème de commande déterministe,
- en ne laissant que deux joueurs, avec en outre $L^2_t = -L^1_t =: L_t$ (et de même pour les K^i_t), et en retirant les aléas, on a l'équation d'Isaacs d'un jeu à deux joueurs et somme nulle, où on peut encore considérer différentes variantes :
  - en faisant systématiquement jouer les deux joueurs simultanément à tous les coups, on cherchera un point selle (particularisation au jeu à deux joueurs et somme nulle de l'équilibre de Cournot-Nash),
  - en faisant systématiquement alterner joueur 1 puis joueur 2, on recherchera une valeur supérieure, tandis que si on fait jouer les joueurs dans l'ordre inverse, on aura une valeur inférieure. Remarquons que ces cas mènent à des décisions "pures" (non randomisées) à chaque pas. Ils sont particulièrement utilisés en commande robuste,
- en ne rendant systématiquement "actif" qu'un seul joueur à chaque coup, et en restreignant tous les ensembles à des ensembles finis, on obtient un arbre de décision classique, menant à des décisions "pures". La version avec plusieurs joueurs actifs simultanément peut être reconstituée avec des ensembles d'information adéquates.

1 Jeu dynamique à somme non nulle

1.1 Description informelle

Un jeu dynamique (à plusieurs joueurs et somme non nulle) est constitué d'une suite de "coups". (Nous ne considérerons ici que des situations où le nombre total de coup est fini, inférieur à un nombre donné.) À chaque coup, un certain nombre de joueurs, dits "actifs" à ce coup là, doivent prendre une décision simultanément, mais en connaissant le résulat de tous les coups précédents. L'ensemble des décisions parmi lesquelles chacun d'entre eux doit choisir peut dépendre du coup. Ces décisions ont des effets de deux ordres :

- d'une part, ils produisent un paiement attaché à ce coup pour tous les joueurs (ou tous les joueurs actifs, comme on voudra), peut-être de façon probabiliste
- d'autre part, ils influencent, de façon déterministe ou probabiliste, ce que sera le coup prochain : joueurs actifs, paiements,....

Les aléas évoqués ici sont *indépendants*, au sens probabiliste, d'un coup à l'autre. Par contre, leur loi de probabilité, supposée connue des joueurs, peut dépendre du coup.

Il est clair que si les coups a priori possibles sont en nombre fini, ainsi que les décisions offertes au choix des joueurs à chaque coup, ceci peut être représenté par un arbre de jeu en forme extensive "à la Kuhn", quitte à exploiter adroitement les ensembles d'information pour représenter des joueurs jouant simultanément.

1.2 Description formelle

On donne,

- $N \in \mathbb{N}$, le nombre de joueurs, $\mathcal{I} = \{1, 2, \ldots, N\}$.
- $T \in \mathbb{N}$ le nombre maximum de coups d'une partie. On notera $\boldsymbol{T} = \{0, 1, \ldots, T-1\}$. On notera $t \in \{0, 1, \ldots, T-1\}$ le numéro d'un coup, et on parlera d'*étape* t. La notation $\forall t$ voudra dire $\forall t \in \boldsymbol{T}$.
- $\forall t$, X_t l'ensemble des *états* possibles à l'étape t. On désigne par $x_t \in X_t$ l'état atteint à l'étape t. (Il définit, comme on va voir les règles du coup à jouer.) La notation $\forall (t, x)$ (ou $\forall (t, x_t)$) voudra dire $\forall t$, $\forall x \in X_t$ (resp. $\forall t$, $\forall x_t \in X_t$).
- Un ensemble $\mathcal{T} \subset (1, X_1) \cup (2, X_2) \cup \cdots \cup (T, X_T)$ d'états dits *terminaux*, qui contient tout (T, X_T). Le jeu s'arrête quand il atteint un état terminal. On notera t_f l'instant où $\mathcal{T}$ est atteint:

  $$(t_f, x_{t_f}) \in \mathcal{T}\,. \tag{1}$$

  En prenant le cas $\mathcal{T} = (T, X_T)$, on a le cas où le nombre d'étapes du jeu est défini a priori, et alors $t_f = T$.
- $\forall (t, x)$, $\forall i \in \mathcal{I}$, $U^i(t, x)$ l'ensemble des actions offertes au joueur i au coup (t, x). On notera $u_t^i \in U^i(t, x_t)$ la décision choisie par le joueur i au coup (t, x_t), et u_t le vecteur des u_t^i.

  Une remarque importante ici est qu'un certain nombre de $U^i(t, x)$ peuvent être réduits à un singleton, dont l'unique objet sera appelé "nill", ou "passe", ou 0,.... C'est à dire que ces joueurs sont inactifs à ce coup là. Il sera commode de désigner par $A(t, x) \subset \mathcal{I}$ l'ensemble

des joueurs actifs au coup (t, x), c'est à dire dont l'ensemble d'alternatives à ce coup n'est pas réduit à un singleton.

- $\forall (t, x)$, l'ensemble $W(t, x)$ des aléas possibles au coup (t, x), et la loi de probabilité $P(t, x)(\cdot)$ régissant $w \in W(t, x)$. Rappelons que les aléas sont supposés indépendants d'une étape à l'autre. (On dit que la suite des $\{w_t\}$ est *blanche*.)

- $\forall t, \forall i \in \mathcal{I}$, les fonctions $L_t^i : X_t \times U^1(t, x) \times \cdots \times U^N(t, x) \times W_t \to \mathbb{R}$, donnant le paiement du joueur i associé au coup de l'étape t si les paramètres en sont ceux donnés en argument,

- $\forall (t, x) \in \mathcal{T}, \forall i \in \mathcal{I}$, les réels $K_t^i(x)$ donnant le paiement au joueur i associé au fait d'atteindre cet état terminal,

- $\forall t$, les fonctions $f_t : X_t \times U^1(t, x) \times \cdots \times U^N(t, x) \times W_t \to X_{t+1}$, qui donnent l'état atteint à l'étape $t + 1$ en fonction de ce qui s'est passé à l'étape t et de l'aléa, par ce que nous appelons l'équation dynamique (ou la description de l'arbre si tous les ensembles sont finis) :

$$x_{t+1} = f_t(x_t, u_t, w_t)\,. \tag{2}$$

1.3 Solutions et stratégies

On cherche non pas des suites $\{u_t^i\}$ "optimales" dans un certain sens, mais de règles de décision qui exploitent l'information sur le passé du jeu, à savoir a priori des règles de la forme

$$u_t^i = \varphi_t^i(x_0, x_1, \ldots, x_t)\,, \tag{3}$$

(ce qu'on appelle des *commandes en boucle fermée*) voire pour des décisions "mixtes" ou randomisées, des lois de probabilité

$$p_t^i = \psi_t^i(x_0, x_1, \ldots, x_t) \tag{4}$$

régissant chacun des u_t^i, supposés probabilistiquement indépendants si nous renonçons (sagement) aux équilibres corrélés. C'est ce qu'on appellera des *stratégies comportementales*.

Étant donné un état initial x_0 et un ensemble de stratégies définies pour tout t et tout i $\{\psi_t^i(\cdot)\}$, qu'on notera simplement ψ, qui engendre un processus $\{x_t\}$, on notera

$$J^i(x_0; \psi) = \mathbb{E}\left(K_{t_f}^i(x_{t_f}) + \sum_{t=0}^{t_f-1} L_t^i(x_t, u_t, w_t)\right) \tag{5}$$

le paiement espéré du joueur i.

Si on a convenu qu'on appellera toujours de noms différents des états provenant d'origines différentes, c'est à dire que l'arbre de jeu est ... un arbre, précisément, alors se souvenir des x_t passés est mainifestement inutile, puisqu'ils sont impliqués par l'état présent. Mais cette convention peut être inutile et excessivement dispendieuse, au point d'être inutilisable. Dans ce cas (où le même (t, x_t) peut être atteint par plusieurs suites de coups différentes, mais de même longueur —ce qui est beaucoup moins contraignant), on insiste sur le fait qu'on veut permettre des décisions de la forme (3). Mais on montrera que les équilibres trouvés sont de la forme

$$p_t^i = \psi_t^i(x_t)\,, \tag{6}$$

apelée *feedback d'état* ou (plus français) *rétroaction d'état*.

On fait remarquer qu'il n'y a pas, dans cette forme, de problème de mémoire de ses propres coups passés. On pourrait imaginer que des coups différents mènent au même état (t, x), on n'a alors pas besoin de se souvenir du chemin suivi. Mais celà n'est pas la même chose que d'avoir, dans un arbre de jeu en forme extensive, un enesmble d'information qui contienne deux nœuds correspondant à des décisions différentes du joueur dans le passé. En effet, ici, ce ne sont pas deux nœuds différents mais indiscernables pour le joueur au moment de prendre sa décision : c'est le même nœud, c'est à dire que toute la suite du jeu est identique.

2 Programmation dynamique

Nous recouvrons de ce terme de *progammation dynamique*, inventé par Bellman, ce que les joueurs appellent "backward induction", ou ce qu'Isaacs appelait (avant Bellman) "tenet of transition".

2.1 Équilibres

Étant donné un jeu défini par N fonctions de N variables ("de décision"), disons $H^i(u)$, $i = 1, 2, \ldots, N$, $u \in U^1 \times U^2 \times \cdots \times U^N$, on appelle

$$\widetilde{\max_u} H(u)$$

le vecteur des valeurs d'un équilibre de Nash. Ces équilibres (optimaux) ne sont pas nécessairement uniques.

On définit alors la procédure de programmation dynamique de la façon suivante. On calcule N fonctions (rélles) $V^i(t, x)$ en progressant des coups terminaux vers le début :

$$\forall (t_f, x_f) \in \mathcal{T}\,, \quad V^i(t_f, x_f) = K^i_{t_f}(x_f) \tag{7}$$

$$\forall (t, x)\,, \quad V^i(t, x) = \widetilde{\max_u}\, \mathbb{E}_w^{P(t,x)}[V^i(t+1, f_t(x, u, w)) + L^i_t(x, u, w)]\,. \tag{8}$$

L'espérance ci-dessus est comprise par rapport à w régi par la loi de probabilité $P(t, x)$, toutes les autres variables figées. (Elle sera notée ci-dessous $\mathbb{E}^{t,x}_{w_t}$.) Les stratégies (mixtes) d'équilibre dans (8) correspondant à l'équilibre retenu seront notées ψ^*_t, de composantes ψ^{i*}_t, $i = 1, \ldots, N$. Elles sont définies en chaque (t, x), et constituent donc ensemble des stratégies comportementales.

On annonce le théorème :

Théorème 1 *Les stratégies d'équilibre ψ^{i*}_t obtenues par la procédure de programmation dynamique (7,8) forment un équilibre de Nash en stratégies comportementales (resp. un point selle, ou des feedbacks d'état optimaux si ce sont des stratégies pures). La fonction $V^i(t, x)$ donne le paiement (optimal) obtenu dans cette solution pour le sous-jeu* commençant *en (t, x), et en particulier, la valeur (optimale) du problème pour le joueur i est $V^i(0, x_0)$.*

2.2 Preuve du théorème

Considérons alors un jeu de stratégies comportementales (de feedbacks d'état) noté ψ^{*-i} où tous les joueurs utilisent leur stratégie d'équilibre calculée dans (8) sauf le joueur i qui adopte une stratégie ψ^i quelconque. Considérons le processus stochastique $\{x_t\}$ engendré depuis x_0 par ces tratégies.

Notons qu'en *tout* (t, x), on a, par la définition d'un équilibre de Nash (ou simplement d'un max)

$$V^i(t, x) \geq \mathbb{E}_u^{\psi^{*-i}} \mathbb{E}^{t,x}_{w_t}[V^i(t+1, f_t(x, u, w)) + L^i_t(x, u, w)]\,.$$

où l'espérance $\mathbb{E}_u^{\psi^{*-i}}$ est prise avec les u^j régis par les lois de probabilitées ψ^{*-i}. En particulier, pour toute réalisation du processus $\{x_t\}$, et $\forall t$, on a

$$V^i(t, x_t) \geq \mathbb{E}_{u_t}^{\psi^{*-i}} \mathbb{E}_{w_t}^{t,x}[V^i(t+1, f_t(x_t, u_t, w)) + L_t^i(x_t, u_t, w)]\,.$$

Un argument probabiliste, que nous repoussons en annexe, permet alors d'affirmer que, *si la stratégie ψ^i est bien causale* (c'est ici que cette condition intervient), on peut prendre les espérances a priori pour en conclure, pour les processus stochastiques $\{x_t\}$ et $\{u_t\}$:

$$0 \geq \mathbb{E}[V^i(t+1, f_t(x_t, u_t, w)) - V^i(t, x_t) + L_t^i(x_t, u_t, w)]\,. \tag{9}$$

Pour nous affranchir du problème du temps d'arrêt t_f, imaginons qu'on a étendu les données en ajoutant à chaque X_t un état "puits" (noté ω), que pour tout $(t_f, x_f) \in \mathcal{T}$, on définit $L_{t_f}^i(x_f, u, w) = K_{t_f}^i(x_f)$ et $f_{t_f}(x_f, u, w) = \omega$, puis que dans les états puits les L^i sont tous nuls et les f_t envoient tous à l'état puits du temps suivant, l'état puits de X_T étant doté d'un $K_T(\omega) = 0$. Ainsi on peut considérer que le jeu procède jusqu'en $t = T$ quoi qu'il arrive. On somme alors la dernière inégalité ci-dessus de $t = 0$ à $t = T - 1$, et il vient

$$0 \geq \mathbb{E}[V^i(T, x_T) - V^i(0, x_0) + \sum_{t=0}^{T-1} L_t^i(x_t, u_t, w_t)]\,,$$

et en utilisant (7) (et la remarque que notre extension a pour résultat que V_t^i est constante de t_f à T)

$$V^i(0, x_0) \geq \mathbb{E}[V^i(t_f, x_{t_f}) + \sum_{t=0}^{t_f-1} L_t^i(x_t, u_t, w_t)]\,,$$

ou encore

$$V^i(0, x_0) \geq J^i(x_0; \psi^{*-i})\,.$$

Refaisons le même calcul, mais avec ψ^* au lieu de ψ^{*-i}. Toutes les inégalités sont remplacées par des inégalités, et on conclue donc qu'alors

$$V^i(0, x_0) = \mathbb{E}[V^i(t_f, x_{t_f}) + \sum_{t=0}^{t_f-1} L_t^i(x_t, u_t, w_t)]\,,$$

ou encore

$$V^i(0, x_0) = J^i(x_0; \psi^*)$$

En comparant ces deux derniers résultats, on obtient le théorème annoncé. (On remarque que dans la version "Kolmogorov", seul le deuxième calcul, avec les égalités, intervient.)

3 Annexe

Le résultat dépend des lemmes suivants :

Lemme 1 *Soit $(\Omega, \mathcal{A}, P)$ un espace probabilisé. Soit $\mathcal{F}$ une sous-σ-algèbre de $\mathcal{A}$. Soit $Y : \Omega \to \mathcal{Y}$ et $Z : \Omega \to \mathcal{Z}$ deux variables aléatoires, Y mesurable sur $\mathcal{F}$ et Z indépendante de $\mathcal{F}$. Soit $g : \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}$ une fonction (rélle) mesurable, et, pour tout $y \in \mathcal{Y}$, posons $h(y) := \mathbb{E}g(y, Z)$. Alors, $\mathbb{E}^{\mathcal{F}}(g(Y, Z)) = h(Y)$.*

Soit $\mathcal{A}_t$ la sigma algèbre engendrée par tous les aléas jusqu'à $t-1$, et $\mathcal{U}_t$ la sigma algèbre dont est doté implicitement le produit cartésien des $U_t^i(x)$ pour définir les stratégies mixtes. Utilisons le premier lemme avec $\mathcal{F}_t = \mathcal{A}_t \otimes \mathcal{U}_t$ pour $\mathcal{F}$, (x_t, u_t) pour Y, w_t pour Z et toute l'expression, $[V^i(t+1, f_t(x_t, u, w)) + L_t^i(x_t, u, w)]$ pour g. Il en découle que l'inégalité précedente peut s'écrire

$$V^i(t, x_t) \geq \mathbb{E}^{\mathcal{F}_t}[V^i(t+1, f_t(x_t, u_t, w)) + L_t^i(x_t, u_t, w)]\,.$$

On obtient (9) en prenant l'espérance a priori des deux membres en utilisant le lemme classique :

Lemme 2 *Pour toute variable aléatoire X et toute sous-sigma algèbre $\mathcal{F}$, on a $\mathbb{E}(\mathbb{E}^{\mathcal{F}} X) = \mathbb{E}X$.*