

**Exercice 10 (Décision dynamique en horizon infini).**

1. Contrairement aux autres exercices déjà traités, l'arbre de jeu est ici infini. Les noeuds de cet arbre sont des suites finies d'états et d'actions. Quand on fixe un état de départ  $k$ , la racine de l'arbre est la suite  $k$ . Le noeud  $(k_1, a_1, k_2, \dots, k_n)$  a pour fils l'ensemble des parties finies  $(k_1, a_1, k_2, \dots, k_n, a_n, k_{n+1})$  telles que  $k_{n+1} = T(k_n, a_n)$ .

Une stratégie est une application  $s : (K \times A)^* \rightarrow (K \times A)$  qui associe à toute suite finie d'états et d'actions  $(k_1, a_1, k_2, \dots, k_n)$  un couple  $(a_n, k_{n+1}) = s(k_1, a_1, \dots, k_n)$  tel que  $k_{n+1} = T(k_n, a_n)$ .

De manière équivalente, on peut se contenter de spécifier quelle action est choisie, et on considèrera qu'une stratégie est une application  $s : (K \times A)^* \rightarrow A$ .

Etant donné une stratégie  $s$  et un état initial  $k$ , il existe un unique chemin infini  $p(k, s) = k_1 a_1 k_2 \dots$  dans l'arbre de jeu tel que  $k_1 = k$  et  $\forall n, a_n = s(k_1, a_1, \dots, k_n)$ .

L'utilité de la stratégie  $s$  quand l'état initial est  $k$  est :

$$U_k(s) = f_\beta(p(k, s))$$

avec

$$f_\beta(k_1 a_1 \dots) = \sum_{n \geq 1} \beta^{n-1} g(k_n, a_n).$$

On en déduit l'égalité :

$$(1) \quad U_k(s) = g(k, s(k)) + \beta \cdot U_{T(k, s(k))}(s[k]),$$

ou  $s[k]$  est la stratégie obtenue par décalage de  $s$  en posant

$$s[k](k_1, a_1, \dots, k_n) = \begin{cases} s(k, s(k), T(k, s(k)), a_1, \dots, k_n) & \text{si } k_1 = T(k, s(k)) \\ s(k_1, a_1, \dots, k_n) & \text{sinon.} \end{cases} .$$

2. Montrons que  $F$  est  $\beta$ -contractante. Comme  $\beta < 1$ , le théorème de point fixe de Picard assurera l'existence d'un unique point fixe. Précisément, montrons que pour tous  $x, y$ ,

$$\|F(x) - F(y)\|_\infty \leq \beta \|x - y\|_\infty .$$

Fixons la composante  $k$ . Soit  $a_{x,k}$  l'élément de  $A$  atteignant l'argmax dans la définition de  $F(x)_k$ . On a

$$\begin{aligned} F(x)_k - F(y)_k &\leq \left( g(k, a_{x,k}) + \beta x_{T(k, a_{x,k})} \right) - \left( g(k, a_{x,k}) + \beta y_{T(k, a_{x,k})} \right) \\ &= \beta \left( x_{T(k, a_{x,k})} - y_{T(k, a_{x,k})} \right) \leq \beta \|x - y\|_\infty . \end{aligned}$$

Par symétrie, notre assertion est donc prouvée.

3. Une stratégie optimale, notée  $s_{opt}$ , consiste à jouer à chaque tour l'argument maximum dans la définition de  $F(v)_k$ . Formellement, soit  $k_1, a_1, \dots, k_n$  une partie finie alors  $s_{opt}(k_1, a_1, \dots, k_n) = a_n$  où

$$a_n = \operatorname{argmax}_{a \in A} \{ g(k_n, a) + \beta v_{T(k_n, a)} \} .$$

Montrons que l'utilité de cette stratégie en partant de l'état  $k$  est bien  $v_k$ .

Remarquons tout d'abord que cette stratégie a la propriété remarquable d'être *positionnelle* : quelque soit le chemin  $k_1 a_1 \dots k_n$  déjà parcouru dans l'arbre, le coup conseillé par  $s_{opt}$

ne dépend que de  $k_n$ . On peut donc la représenter de manière équivalente comme un objet fini  $s_{opt} : K \rightarrow A$ .

Considérons l'équation (1) appliquée à  $s = s_{opt}$ . Comme  $s_{opt}$  est positionnelle, on en déduit que pour tout état  $k$ , les stratégies  $s[k]$  et  $s$  sont identiques. On en déduit que pour tout  $k$ ,  $a = s_{opt}(k)$  et  $l = T(k, a)$ ,

$$U_k(s_{opt}) = g(k, a) + \beta \cdot U_l(s_{opt}) .$$

D'autre part, par définition de  $s_{opt}$ , on a également :

$$v_k = g(k, a) + \beta \cdot v_l .$$

Ainsi,  $U(s_{opt})$  et  $v$  sont tout deux points fixes de l'opérateur  $G : \mathbb{R}^K \rightarrow \mathbb{R}^K$  défini par :

$$G(x)_k = g(k, s_{opt}(k)) + \beta \cdot x_{T(k, s_{opt}(k))} .$$

Or l'opérateur  $G$  est, comme  $F$ , lui aussi  $\beta$ -contractant. On en déduit donc par unicité du point fixe que  $U(s_{opt}) = v$ .

**4.** Montrons que  $W$  est un point fixe de l'opérateur  $F$ .

$$\begin{aligned} W(k) &= \sup_s U_k(s) && \text{par def de } W \\ &= \sup_s (g(k, s(k)) + \beta \cdot U_{T(k, s(k))}(s[k])) && \text{par (1)} \\ &= \sup_{a, s \text{ t.q. } s(k)=a} (g(k, a) + \beta \cdot U_{T(k, a)}(s[k])) \\ &= \sup_a (g(k, a) + \beta \cdot \sup_{s \text{ t.q. } s(k)=a} U_{T(k, a)}(s[k])) \\ &= \sup_a (g(k, a) + \beta \cdot \sup_s U_{T(k, a)}(s)) \\ &= F(W)_k \end{aligned}$$

L'avant-dernière égalité provient du fait que pour toute action  $a$ , l'ensemble  $\{s[k] \mid s(k) = a\}$  est en fait l'ensemble de toute les stratégies.

Par unicité du point fixe de l'opérateur contractant  $F$ , on en déduit que  $v = W = U(s_{opt})$ . Par définition de  $W$ , on a pour tout état  $k$ ,

$$U_k(s_{opt}) = \sup_s U_k(s).$$

En d'autre termes, la stratégie  $s_{opt}$  est optimale, quelque soit l'état initial. On a ainsi montré l'existence d'une stratégie à la fois optimale et positionnelle.

**5.** (non indispensable)

L'ensemble des stratégies peut être muni d'une structure d'espace métrique : la distance entre une stratégie  $s$  et une stratégie  $t$  est  $2^{-l}$ , en notant  $l$  la longueur de la plus petite suite  $k_1 a_1 \cdots k_l$  d'états et d'actions telle que  $s(k_1 a_1 \cdots k_l) \neq t(k_1 a_1 \cdots k_l)$ .

On vérifie que muni de cette métrique, l'ensemble des stratégies est un espace compact.

Avec cette métrique, l'application  $(k, s) \mapsto U_k(s)$  qui associe à une stratégie  $s$  et un état  $k$  la valeur  $U_k(s)$  est continue.

Comme une fonction continue à valeur réelle atteint son maximum sur un compact, on en déduit l'existence d'une stratégie optimale. C'est un résultat moins fort que celui de la question 4, où l'on a déjà prouvé l'existence d'une stratégie optimale, qui est de plus positionnelle.

**6.** La solution est plus amusante si on considère qu'un joueur Joyeux qui se repose gagne 6.

Pour calculer les valeurs du jeu, il suffit d'après la question 4 de calculer les valeurs associées à chaque stratégie positionnelle, et de sélectionner la meilleure. Il y a quatre stratégies positionnelles possibles.

On obtient le tableau suivant, qui indique le paiement reçu en fonction de l'état initial et de la stratégie positionnelle choisie.

	$J \rightarrow R, D \rightarrow R$	$J \rightarrow R, D \rightarrow T$	$J \rightarrow T, D \rightarrow R$	$J \rightarrow T, D \rightarrow T$
$(1 - \beta)U_J(s)$	1	1	$\frac{10}{1+\beta}$	$10 - 8\beta$
$(1 - \beta)U_D(s)$	$\beta$	2	$\beta \frac{10}{1+\beta}$	2

Par exemple, pour la stratégie  $J \rightarrow T, D \rightarrow R$ , quand on part de l'état  $J$  alors la suite des valeurs vues est  $10, 0, 10, 0, \dots$  donc le paiement reçu est :

$$U_J(s) = 10 + \beta \cdot 0 + \beta^2 \cdot 10 + \beta^3 \cdot 0 + \dots = \frac{10}{(1 - \beta)^2}.$$

Une petite étude de la fonction  $\max_s (1 - \beta)U_J(s)$ , où le max est pris sur les quatre stratégies positionnelles, permet d'isoler les trois comportements suivants.

Si  $\beta$  est compris entre 0 et  $\frac{1}{4}$  alors le joueur joue à court terme : les premiers instants du jeu ont beaucoup de poids dans le paiement total. Le joueur choisit de toujours travailler, et est déprimé en permanence.

Si  $\beta$  est compris entre  $\frac{2}{3}$  et 1 alors le joueur joue à très long terme : les premiers instants de la partie ont peu d'importance. Le joueur choisit de toujours se reposer et est toujours joyeux.

Si  $\beta$  est compris entre  $\frac{1}{4}$  et  $\frac{2}{3}$  alors on est dans un cas intermédiaire : quand le joueur est joyeux, il est plus attiré par le gain de 10 qu'il peut obtenir en travaillant que par le gain de 6 qu'il obtient en se reposant. Inversement, quand le joueur est déprimé, il accepte d'encaisser un paiement de 0, car ce faible revenu sera compensé au tour suivant. Finalement, son comportement optimal consiste à alterner une phase de travail et une phase de repos...