Conditions générales pour l'admissibilité de la programmation dynamique dans la décision séquentielle possibiliste

Paul Weng

LIP6
104 avenue du Président Kennedy
75016 Paris
paul.weng@lip6.fr

RÉSUMÉ. *Nous nous intéressons à la contrepartie possibiliste des processus de décision markoviens. À l'instar du modèle classique, trois relations de préférence peuvent être distinguées (préférences sur les chemins, sur les loteries et sur les politiques). Nous énonçons des propriétés simples et suffisantes (transitivité, invariance par translation, indépendance) sur la relation de préférence sur les loteries pour permettre l'utilisation de méthodes fondées sur la programmation dynamique. Nous fournissons enfin un exemple d'application de ces résultats avec l'utilité bipolaire possibiliste.*

ABSTRACT. *We are interested here in the possibilistic counterpart of Markov decision processes. Like in the standard model, three preference relations can be distinguished (preferences over paths, over lotteries and over policies). We state some simple and sufficient properties (transitivity, invariance by translation, independence) on the preference relation over lotteries to allow the application of techniques based on dynamic programming. Finally we provide an example illustrating the interest of these results with binary possibilistic utility.*

MOTS-CLÉS : *processus de décision markoviens, théorie des possibilités, préférence qualitative, programmation dynamique.*

KEYWORDS: *Markov decision process, possibility theory, qualitative preference, dynamic programming.*

RSTI - RIA – 21/2007. Prise de décision séquentielle, pages 129 à 143

1. Introduction

Le modèle des processus de décision markoviens (PDM) est le modèle standard pour la résolution des problèmes de planification dans l'incertain. Il nécessite que l'incertain soit modélisé par la théorie des probabilités et que les récompenses soient numériques et additives. Par conséquent, l'utilisation de ce modèle suppose de préalablement quantifier les probabilités et les récompenses du problème. Cependant dans certaines situations, il n'est pas possible ou difficile de déterminer de manière précise ces valeurs, comme cela est le cas dans les problèmes où le décideur ne connaît son environnement que de manière partielle ou imparfaite. Dans ces cas de figure, il peut alors être souhaitable de recourir à d'autres types de représentation pour l'incertain et les récompenses.

Nous nous intéressons ici au cas où l'incertain est possibiliste et les récompenses sont qualitatives et étudions la contrepartie possibiliste des processus de décision markoviens dans le but de délimiter une large classe de problèmes résolubles par les techniques de programmation dynamique. Dans cet article, nous nous restreignons aux problèmes à horizon fini.

Les PDM possibilistes ont été étudiés par Sabbadin (Sabbadin, 1998, Sabbadin *et al.*, 1998, Sabbadin, 1999) dans le cas où les utilités optimistes ou pessimistes (Dubois *et al.*, 1998, Dubois *et al.*, 2001) sont utilisées. Les méthodes fondées sur la recherche arrière permettent alors de déterminer des politiques optimales. À la différence des travaux sur les PDM possibilistes, nous n'étudions pas un critère en particulier, mais recherchons des conditions générales suffisantes pour garantir l'application de méthodes fondées sur la recherche arrière. Ainsi, ce travail fait suite à celui de Weng (Weng, 2006) dans lequel il est montré dans le cadre classique des PDM que la transitivité, l'indépendance et l'invariance par translation de la relation de préférence sur les loteries permettent l'utilisation de techniques de programmation dynamique.

Le reste de l'article est organisé de la manière suivante. Dans la section 2, nous présentons le modèle des PDM possibilistes généralisés (PDMG), les notations et les définitions utilisées. Ensuite dans la section 3, nous explicitons les trois relations de préférence définies (sur les historiques, sur les loteries et sur les politiques) dans un PDMPG et donnons les conditions suffisantes pour obtenir la propriété de stabilité permettant le fonctionnement de la programmation dynamique. Dans la section 4, nous rappelons certaines des propositions obtenues dans (Weng, 2006) qui s'appliquent également dans ce cadre. Quand la relation de préférence sur les loteries est transitive, indépendante et invariante par translation (cadre des préférences partielles), l'algorithme de recherche arrière généralisé (alg. 4.1) permet le calcul itératif de politiques préférées. Si, de plus, la relation est complète, une version simplifiée (alg. 4.2) de l'algorithme 4.1 est fournie. Enfin un exemple d'application de ces résultats est proposé dans la section 5 avec l'utilité bipolaire possibiliste proposée par Giang et Shenoy (Giang *et al.*, 2001).

2. Cadre général de l'étude

2.1. *Processus de décision markoviens possibilistes généralisés*

Nous supposons que l'incertain est modélisé par des distributions de possibilité (également appelées loteries) et qu'il est mesuré sur un sous-ensemble fini $\mathbf{L}$ de $[0, 1]$. Le plus grand élément de $\mathbf{L}$ est égal à 1 et le plus petit élément à 0. La relation d'ordre sur $\mathbf{L}$ est notée $\geq$. Les opérateurs $\max$ et $\min$ sur $\mathbf{L}$ sont notés respectivement $\vee$ et $\wedge$. Une loterie sur un ensemble X est notée $[\lambda_1/x_1, \ldots, \lambda_n/x_n]$ avec $\forall i = 1, \ldots, n, x_i \in X, \lambda_i \in \mathbf{L}$ et $\bigvee_{i=1}^{n} \lambda_i = 1$. Classiquement, la réduction de loteries composées, c'est-à-dire de loteries définies sur des loteries est définie par l'égalité suivante : $[\lambda/\pi, \mu/\pi'](x) = (\lambda \wedge \pi(x)) \vee (\mu \wedge \pi'(x))$ où $\lambda, \mu \in \mathbf{L}, \lambda \vee \mu = 1$ et π, π' sont deux loteries. Cette formule se généralise naturellement à un nombre quelconque de loteries.

Le modèle des processus de décision markoviens possibilistes généralisés (PDMPG) est défini par la donnée du quadruplet (S, A, T, R) :

1) S l'ensemble fini des états,

2) A l'ensemble fini des actions,

3) $T : S \times A \rightarrow \Pi(S)$ la fonction de transition où $\Pi(S)$ est l'ensemble des distributions de possibilité sur S,

4) $R : S \times A \times S \rightarrow (X, \circ, \succeq)$ la fonction de récompense où X est l'ensemble de valuation des récompenses.

L'ensemble des récompenses X est muni d'un opérateur interne $\circ$ et d'une relation d'ordre $\succeq$. Pour la loi de composition interne $\circ$ définie sur X, on définit pour tout couple $(x, z) \in X \times X$, l'ensemble noté $z \bullet x = \{y \in X \mid x \circ y = z\}$. Cet ensemble peut évidemment être vide. Quand $(X, \circ, \succeq) = (\mathbb{R}, +, \geq)$, on a alors $z \bullet x = \{z - x\}$. Remarquons que $x \circ (z \bullet x) = \{z\}$ quand $z \bullet x \neq \emptyset$.

Les historiques dans ce modèle, débutant dans l'état s, correspondent aux séquences suivantes :

$$(s, a_1, s_1, a_2, s_2, \ldots) \text{ où } \forall i \geq 1, (a_i, s_i) \in A \times S.$$

La valeur d'un historique $\gamma = (s_0, a_1, s_1, a_2, s_2, \ldots, a_n, s_n)$ vaut $x = x_1 \circ \ldots \circ x_n \in X$ où $\forall i = 1, \ldots, n, x_i = R(s_{i-1}, a_i, s_i)$. L'opérateur $\circ$ est supposé associatif. Cette hypothèse permet l'évaluation itérative d'un historique. La structure $(X, \circ, \succeq)$ est choisie de telle sorte qu'elle représente les préférences sur les historiques. La relation $\succeq$ de l'ensemble X correspond donc à la relation de préférence sur les historiques.

Une règle de décision est une fonction de l'ensemble des états S dans l'ensemble des actions A. L'ensemble des règles de décision sera noté $\Delta = A^S$. Une politique à un horizon n est une séquence de n règles de décision. L'ensemble des politiques à l'horizon n sera noté Φ_n. Si $\phi_n \in \Phi_n$, on a alors $\phi_n = (\delta_1, \ldots, \delta_n)$ où chaque $\delta_i \in \Delta$. La politique à l'horizon 0, ne contenant aucune règle de décision, est notée

$()$. Pour une politique ϕ et une règle de décision δ, on note (δ, ϕ) la politique qui consiste à appliquer la règle de décision δ à l'étape 1 et à utiliser la politique ϕ ensuite. Par extension, on écrit (a, ϕ) la règle applicable dans un état, qui consiste à exécuter l'action a dans cet état puis la politique ϕ. Enfin pour un ensemble de politiques Φ, on note $(a, \Phi) = \{(a, \phi) \mid \phi \in \Phi\}$.

Remarquons qu'une règle de décision δ pour un état s définit une loterie sur l'ensemble X. Cette loterie est égale à $[T(s, \delta(s), s')/R(s, \delta(s), s']_{s' \in S}$. Par conséquent, une politique ϕ_n induit, pour un horizon n fixé et un état initial s donné, une loterie sur X également. Nous noterons $L_s^{\phi_n}$ la loterie sur l'ensemble de valuations X induite par la politique ϕ_n à l'état s. Elle associe à tout $x \in X$ la possibilité :

$$L_s^{\phi_n}(x) = \bigvee_{s' \in S} T(s, \delta(s))(s') \wedge L_{s'}^{\phi_{n-1}}(x \bullet R(s, \delta(s), s'))$$

où $\phi_n = (\delta, \phi_{n-1})$ et $\delta \in \Delta$, $\phi_{n-1} \in \Phi_{n-1}$.

Il est donc possible d'étudier ce modèle selon les propriétés de cet ensemble X. On constate que si l'on prend $(X, \circ, \succeq) = (\mathbf{L}, \wedge, \geq)$, on retrouve les PDM possibilistes définis par Sabbadin (Sabbadin, 1998). Si l'on prend $(X, \circ, \succeq) = (\mathbf{L}^p, \wedge, \geq_D)$ pour $p > 0$, on obtiendrait alors le modèle (non encore étudié) des PDM possibilistes multicritères avec la relation de dominance de Pareto $\geq_D$. Avec $X = S \times A \times S$, le PDMPG correspond au modèle de Sobel (Sobel, 1975) formulé dans l'incertain possibiliste.

2.2. *Définitions et notations*

Pour une relation de préférence $\succsim$, on écrira $\succ$ pour la partie asymétrique et $\sim$ pour la partie symétrique avec leurs sens habituels. La relation $\succsim$ s'interprète comme "au moins aussi bon que", $\succ$ comme "strictement meilleur" et $\sim$ comme " de même qualité". Pour une relation d'ordre $\succeq$, on écrira $\succ$ pour la partie asymétrique et $=$ pour la partie symétrique.

Pour un ensemble Y et une relation de préférence $\succsim$ sur cet ensemble, on définit l'ensemble des éléments maximaux par $M(Y, \succsim) = \{y \in Y \mid \forall z \in Y, \neg(z \succ y)\}$. Quand il n'y a pas d'ambiguïté possible sur la relation de préférence utilisée, on notera simplement cet ensemble $M(Y)$. Si la relation de préférence sur Y est complète, $M(Y)$ est noté $\max(Y)$ et devient simplement l'ensemble des éléments optimaux définis par $\max(Y) = \{y^* \in Y \mid \forall y \in Y, y^* \succsim y\}$.

Si l'on note la relation de préférence sur les politiques $\succsim_\Phi$, alors l'ensemble des politiques maximales ou optimales pour un horizon n donné est noté

$$\Phi_n^* = M(\Phi_n, \succsim_\Phi).$$

De plus, on définit $\forall n > 0, \Phi_n^+$ par

$$\begin{aligned} \Phi_1^+ &= \Phi_1^* \\ \forall n \geq 1, \Phi_{n+1}^+ &= \bigcup_{\phi_n \in \Phi_n^+} M(\{(\delta, \phi_n) \mid \delta \in \Delta\}, \succsim_\Phi). \end{aligned}$$

On remarquera que l'algorithme de recherche arrière construit exactement ces ensembles. Pour chaque politique calculée à l'étape précédente, on calcule la ou les meilleures (au sens de $\succsim_\Phi$) règles de décision à lui ajouter à la première étape.

Enfin, on définit $\forall n > 0, \Phi_n^{+M}$ par

$$\begin{aligned} \Phi_1^{+M} &= \Phi_1^* \\ \forall n \geq 1, \Phi_{n+1}^{+M} &= M(\bigcup_{\phi_n \in \Phi_n^{+M}} \{(\delta, \phi_n) \mid \delta \in \Delta\}, \succsim_\Phi). \end{aligned}$$

Ces ensembles sont également définis de manière récursive. Pour une étape donnée, on considère dans cette définition les meilleures politiques parmi l'ensemble des politiques déterminées précédemment auxquelles on a adjoint une règle de décision. La différence avec la définition précédente est la portée de l'opérateur de maximisation. Pour déterminer un élément de Φ_{n+1}^{+M}, il est nécessaire de calculer entièrement Φ_n^{+M}. Par contre, pour obtenir un élément de Φ_{n+1}^+, il suffit de déterminer un seul élément de Φ_n^+. Notons une réécriture intéressante de Φ_n^{+M} qui nous servira dans la définition des algorithmes :

$$\forall n \geq 1, \Phi_{n+1}^{+M} = M(\bigcup_{\phi_n \in \Phi_n^{+M}} M(\{(\delta, \phi_n) \mid \delta \in \Delta\}, \succsim_\Phi), \succsim_\Phi).$$

Enfin, sous certaines hypothèses (voir prop. 4.5), les deux dernières définitions sont équivalentes.

La propriété suivante d'invariance par translation permet d'affirmer qu'une préférence entre deux loteries est conservée même si tous les éléments sur lesquels sont définies les loteries sont translatés d'une même "quantité". Nous notons $\mathbf{L}(X)$ l'ensemble des loteries possibilistes sur X.

Définition 2.1. *Une relation de préférence $\succsim_L$ sur les loteries définies sur $(X, \circ)$ est* invariante par translation *si et seulement si*

$\forall (L_1, L_2) \in \Pi(X) \times \Pi(X), (L_1 \succsim_L L_2 \Rightarrow \forall r \in X, L_1^{\rightarrow r} \succsim_L L_2^{\rightarrow r})$ *où*

$\forall i = 1, 2, \forall x \in X, L_i^{\rightarrow r}(x) = L_i(x \bullet r)$.

Dans le cadre des PDMPG, l'invariance par translation dit simplement que la préférence entre deux actions ne s'inverse pas si toutes leurs récompenses sont modifiées d'une même valeur.

La propriété d'indépendance que nous introduisons maintenant correspond en fait à une version affaiblie de la propriété d'indépendance de l'axiomatique de von Neumann et Morgenstern (von Neumann *et al.*, 1944) formulée par Fishburn (Fishburn, 1970). Elle dit en substance que les préférences sur deux loteries ne peuvent s'inverser si on combine ces deux loteries à une troisième loterie, c'est-à-dire, de manière intuitive que l'"ajout" de conséquences identiques (avec les mêmes possibilités) à deux loteries ne peut inverser le sens de préférence.

Définition 2.2. *Une relation de préférence $\succsim_L$ sur les loteries vérifie la propriété d'*indépendance *si et seulement si*

$\forall L_1, L_2 \in \Pi(X), (L_1 \succsim_L L_2 \Rightarrow \forall \lambda, \mu \in \mathbf{L}$, *tels que* $\lambda \vee \mu = 1, \forall L_3 \in \Pi(X), [\lambda/L_1, \mu/L_3] \succsim_L [\lambda/L_2, \mu/L_3])$.

L'interprétation de l'indépendance est assez simple dans le cadre des PDMPG. Elle stipule que dans l'application d'une politique dans un état, le fait de remplacer une de ses sous-politiques par une sous-politique qui lui est préférée permet d'obtenir une nouvelle politique au moins aussi bonne. Une forme faible de cette propriété est utilisée dans l'axiomatisation des utilités optimistes et pessimistes (Dubois *et al.*, 1995, Dubois *et al.*, 1998) et il est aisé de montrer que ces utilités vérifient notre propriété d'indépendance.

Nous définissons également la propriété de stabilité sur la relation de préférence sur les politiques. Intuitivement, elle signifie simplement que si une politique ϕ est préférée à une politique ϕ' alors le fait de retarder l'application de ces deux politiques par l'utilisation d'une même règle de décision δ conserve le sens de la préférence. Cette propriété est cruciale pour permettre le calcul itératif de politiques préférées.

Définition 2.3. *Une relation de préférence $\succsim_\Phi$ sur les politiques sera dite* stable *si et seulement si*

$\forall (\phi, \phi') \in \Phi \times \Phi, (\phi \succsim_\Phi \phi' \Rightarrow \forall \delta \in \Delta, (\delta, \phi) \succsim_\Phi (\delta, \phi'))$.

Cette propriété impose une certaine invariance des préférences dans le temps. En effet, si une politique est préférée à une autre à l'instant courant, cette préférence restera vraie à l'étape suivante.

Considérons pour $\delta \in \Delta$, l'opérateur $H_\delta : \Phi \rightarrow \Phi$ qui associe à toute politique ϕ la nouvelle politique (δ, ϕ). Alors la stabilité sur la relation de préférence sur les politiques correspond à la notion de monotonie de l'opérateur H_δ pour toute règle de décision δ.

3. Relations de préférence et stabilité

Comme dans les PDM classiques, il est possible de distinguer trois niveaux de relations de préférence dans le modèle des PDMPGs. Une première relation $\succsim$ est définie sur les historiques. Une politique pour un horizon fixé et un état initial donné

induisant une loterie sur l'ensemble X, comparer deux politiques à un horizon donné et dans un certain état initial revient à comparer leurs loteries respectives. À partir de la première relation de préférence, une relation de préférence $\succsim_L$ sur les loteries est donc définie. Enfin, cette dernière induit une troisième relation de préférence $\succsim_\Phi$ sur les politiques permettant de définir la notion d'optimalité ou de maximalité sur l'ensemble des politiques. La relation $\succsim_\Phi$ est définie par :

$$\forall(\phi,\phi') \in \Phi \times \Phi, \phi \succsim_\Phi \phi' \Leftrightarrow \forall s \in S, L_s^\phi \succsim_L L_s^{\phi'}. \quad [1]$$

Pour clarifier les choses, nous donnons deux exemples. Dans les PDM classiques, la relation de préférence sur les historiques est simplement celle définie par $(\mathbb{R}, \geq)$. En effet, les historiques peuvent être comparés entre eux par la somme des récompenses qu'ils induisent. La relation de préférence sur les loteries est représentée par l'utilité espérée. En effet, à une loterie est associée l'espérance des récompenses et les loteries sont comparées entre elles via ces espérances. Enfin, la relation de préférence sur les politiques est celle définie ci-dessus (eq. [1]).

Dans les PDM possibilistes (Dubois *et al.*, 1996, Sabbadin, 1998, Sabbadin *et al.*, 1998, Sabbadin, 1999), la valeur d'un historique est déterminée par le minimum des récompenses obtenues. La relation de préférence sur les historiques est donc celle de $(L, \geq)$ où L est l'échelle qualitative sur laquelle sont mesurées les récompenses. La relation de préférence sur les loteries est représentée par les utilités optimistes ou pessimistes. Enfin, la relation de préférence sur les politiques se définit classiquement par l'équation [1].

Nous nous intéresserons ici plus particulièrement à la relation de préférence sur les loteries. Nous fournissons des propriétés suffisantes sur cette relation pour garantir la stabilité de la relation de préférence sur les politiques. Dans ce but, nous énonçons d'abord le lemme suivant qui indique que sous les conditions d'indépendance et de transitivité de la relation de préférence sur les loteries la combinaison d'un nombre quelconque de loteries conserve le sens de préférence. Autrement dit, dans notre contexte, ce lemme donne les conditions pour garantir la non-inversion du sens de préférence de deux politiques quand on reporte leurs applications d'une étape par l'utilisation d'une même action à la première étape.

Lemme 3.1. *Si une relation de préférence $\succsim_L$ sur les loteries est indépendante et transitive alors, si $(L_i)_{i=1..n}$ et $(L'_i)_{i=1..n}$ représentent deux familles finies de loteries telles que $\forall i = 1..n, L_i \succsim_L L'_i$, on a $\forall(\lambda_i)_{i=1..n} \in [0,1]$, tels que $\bigvee_{i=1}^n \lambda_i = 1$,*

$$[\lambda_1/L_1, \ldots, \lambda_n/L_n] \succsim_L [\lambda_1/L'_1, \ldots, \lambda_n/L'_n].$$

Démonstration. La démonstration se fait par récurrence sur n.

Pour $n = 2$, prenons deux couples de loteries (L_1, L_2) et (L'_1, L'_2) telles que $L_1 \succsim_L L'_1$ et $L_2 \succsim_L L'_2$. En appliquant la propriété d'indépendance sur la première relation et L_2, on a $\forall \lambda, \mu \in \mathbf{L}$, tels que $\lambda \vee \mu = 1$, $[\lambda/L_1, \mu/L_2] \succsim_L [\lambda/L'_1, \mu/L_2]$. Puis en appliquant la propriété d'indépendance sur la seconde relation et L'_1, on a

$\forall \lambda, \mu \in \mathbf{L},$ tels que $\lambda \vee \mu = 1, [\lambda/L'_1, \mu/L_2] \succsim_L [\lambda/L'_1, \mu/L'_2]$. Enfin par transitivité, on obtient bien : $\forall \lambda, \mu \in \mathbf{L},$ tels que $\lambda \vee \mu = 1, [\lambda/L_1, \mu/L_2] \succsim_L [\lambda/L'_1, \mu/L'_2]$.

Supposons maintenant que la relation est vraie avec n loteries. Considérons deux familles de loteries $(L_i)_{i=1..n+1}, (L'_i)_{i=1..n+1}$ telles que $\forall i = 1..n+1, L_i \succsim_L L'_i$. Soit une séquence $(\lambda_i)_{i=1..n+1} \in \mathbf{L}$ telle que $\bigvee_{i=1..n+1} \lambda_i = 1$.

Cas 1 : $\lambda_{n+1} \neq 1$: Posons $L = [\lambda_1/L_1, \ldots, \lambda_n/L_n]$ et $L' = [\lambda_1/L'_1, \ldots, \lambda_n/L'_n]$. Ce sont deux loteries. Et d'après l'hypothèse de récurrence, $L \succsim_L L'$.

En appliquant la propriété démontrée pour $n = 2$, en prenant $\lambda = 1$ et $\mu = \lambda_{n+1}$, on obtient :

$[1/L, \lambda_{n+1}/L_{n+1}] \succsim_L [1/L', \lambda_{n+1}/L'_{n+1}]$.

En développant L et L', on obtient bien :

$$[\lambda_1/L_1, \ldots, \lambda_{n+1}/L_{n+1}] \succsim_L [\lambda_1/L'_1, \ldots, \lambda_{n+1}/L'_{n+1}].$$

Cas 2 : $\lambda_{n+1} = 1$: On peut faire de même que dans le cas 1 avec λ_1. □

Ce lemme nous permet de prouver la proposition suivante qui fournit des conditions suffisantes pour garantir la stabilité de la relation de préférence sur les politiques.

Proposition 3.1. *Si $\succsim_L$ (resp. $\succ_L$) est transitive, invariante par translation et indépendante alors $\succsim_\Phi$ (resp. $\succ_\Phi$) est stable.*

Démonstration. Soient deux politiques ϕ, ϕ' telles que $\phi \succsim_\Phi \phi'$. Soit une règle de décision δ. Par hypothèse, on a $\forall s' \in S, L^{\phi}_{s'} \succsim_L L^{\phi'}_{s'}$.

Considérons un état initial s quelconque. Par définition, la loterie induite par (δ, ϕ) en s vaut :

$$\forall x \in X, L^{(\delta,\phi)}_s(x) = \bigvee_{s' \in S} T(s, \delta(s))(s') \wedge L^{\phi}_{s'}(x \bullet R(s, \delta(s), s')).$$

De même, pour (δ, ϕ'), on obtient :

$$\forall x \in X, L^{(\delta,\phi')}_s(x) = \bigvee_{s' \in S} T(s, \delta(s))(s') \wedge L^{\phi'}_{s'}(x \bullet R(s, \delta(s), s')).$$

En posant $\forall s' \in S, \forall x \in X, L_{s'}(x) = L^{\phi}_{s'}(x \bullet R(s, \delta(s), s'))$ et $L'_{s'}(x) = L^{\phi'}_{s'}(x \bullet R(s, \delta(s), s'))$, on peut réécrire les loteries $L^{(\delta,\phi)}_s = \bigvee_{s' \in S} T(s, \delta(s))(s') \wedge L_{s'}$ et $L^{(\delta,\phi')}_s = \bigvee_{s' \in S} T(s, \delta(s))(s') \wedge L'_{s'}$. En vertu de l'hypothèse d'invariance par translation, $\forall s' \in S, L_{s'} \succsim_L L'_{s'}$. D'après le lemme précédent 3.1, $\bigvee_{s' \in S} T(s, \delta(s))(s') \wedge L_{s'} \succsim_L \bigvee_{s' \in S} T(s, \delta(s))(s') \wedge L'_{s'}$. On a bien $L^{(\delta,\phi)}_s \succsim_L L^{(\delta,\phi')}_s$. Par conséquent, $\succsim_\Phi$ est stable.

De manière similaire, on démontre que si $\succ_L$ est transitive, invariante par translation et indépendante alors la relation $\succ_\Phi$ associée est stable. □

4. Étude de deux structures de préférence

Dans cette section, nous rappelons les résultats obtenus par Weng (Weng, 2006) d'abord dans le cadre général (préférences partielles) garantissant que des politiques préférées existent et peuvent être construites itérativement par recherche arrière (algo. 4.1) puis dans le cas particulier des préférences complètes. Nous rappelons également le lien entre les deux résultats et fournissons une spécification (algo. 4.2) plus efficace de l'algorithme général précédent.

4.1. *Cadre des préférences partielles*

Le cadre des préférences partielles se définit par la donnée d'une relation de préférence transitive sur les loteries et d'une relation de préférence stable sur les politiques. Il inclurait par exemple le modèle des PDM possibilistes multicritères (non encore étudié). Sous ces conditions, nous démontrons qu'il existe au moins une politique maximale et que l'algorithme 4.1 permet de la calculer itérativement.

Si la relation de préférence sur les loteries est transitive et celle sur les politiques est stable alors une politique maximale existe et il est possible de la construire itérativement, c'est-à-dire, sous ces conditions, l'algorithme de recherche arrière permet le calcul d'un sous-ensemble des politiques maximales.

Proposition 4.1. *Si $\succsim_L$ est transitive et $\succsim_\Phi$ est stable alors pour tout $n > 0$, les ensembles Φ^*_n, Φ^{+M}_n ne sont pas vides et $\Phi^{+M}_n \subseteq \Phi^*_n$.*

Si, de plus, la relation de préférence stricte sur les politiques est stable, la proposition suivante garantit que toute sous-politique d'une politique maximale est maximale. Autrement dit, sous cette dernière condition, toutes les politiques préférées se calculent de manière itérative.

Proposition 4.2. *Si $\succsim_L$ est transitive et les relations $\succsim_\Phi$ et $\succ_\Phi$ sont stables alors pour tout $n > 0$, Φ^*_n n'est pas vide et $\Phi^{+M}_n = \Phi^*_n$.*

De ces propositions, il est possible de définir l'algorithme de recherche arrière généralisé :

1: $t \leftarrow N$
2: $\Phi^{+M}_N \leftarrow \{()\}$
3: **repeat**
4: $\quad t \leftarrow t - 1$
5: $\quad$ **for all** $\phi \in \Phi^{+M}_{t+1}$ **do**
6: $\quad\quad$ **for all** $s \in S$ **do**

7: $\quad\quad \Phi_t^{+M}(s) \leftarrow M(\{(a,\phi) : a \in A\}), \succsim_L$
8: **end for**
9: ajout dans Φ_t^{+M} des politiques obtenues à partir de $\Phi_t^{+M}(s)$
10: **end for**
11: $\Phi_t^{+M} \leftarrow M(\Phi_t^{+M}, \succsim_\Phi)$
12: **until** $t = 0$

Dans chaque état, l'algorithme calcule les actions maximales à effectuer pour l'horizon t (ligne 7). Puis, il construit la ou les meilleures règles de décision pour l'horizon t (ligne 9) en sélectionnant une action parmi la ou les meilleures actions calculées dans chaque état. Ces opérations sont effectuées pour chaque politique maximale calculée à l'étape précédente. Finalement, seules les politiques non dominées sont conservées (ligne 11). L'algorithme calcule donc pour chaque étape Φ_t^{+M}. La propriété $\forall t > 0, \Phi_t^{+M} \subseteq \Phi_t^*$ de la proposition 4.1 garantit que les politiques ainsi déterminées sont maximales. Quand les deux ensembles sont égaux, l'algorithme permet l'obtention de toutes les politiques maximales. Dans cet algorithme, comme il a été signalé lors de la définition de Φ_t^{+M}, même pour obtenir une seule politique maximale à un horizon N, il est nécessaire de calculer tous les éléments de Φ_t^{+M} aux horizons $t < N$.

L'algorithme proposé travaille directement sur l'espace des loteries et utilise les loteries pour comparer les actions. Il est donc très général et peut s'instancier sur différentes structures de préférence (qualitatives notamment) vérifiant les hypothèses de la proposition 4.1. Bien entendu, l'algorithme serait difficilement exploitable directement puisqu'il nécessite le calcul à chaque étape de l'ensemble des récompenses qu'une politique donnée peut générer et les possibilités associées à celles-ci. Dans la pratique, il est nécessaire d'expliciter la relation de préférence sur les loteries et d'utiliser si possible ses propriétés. Par exemple, si la relation est représentable par un critère simple (utilités optimiste, pessimiste ou bipolaire possibiliste), l'algorithme proposé se simplifie naturellement (lignes 7 et 11).

4.2. *Le cadre des préférences complètes*

Le cadre des préférences complètes se définit par la donnée d'une relation de préférence complète et transitive sur les loteries et d'une relation de préférence stable sur les politiques. On constate que la complétude de la relation de préférence sur les loteries est ajoutée aux hypothèses précédentes. La contrepartie possibiliste des PDM (Dubois *et al.*, 1996, Sabbadin, 1998, Sabbadin *et al.*, 1998, Sabbadin, 1999) est un exemple appartenant à cette classe de préférence.

Les résultats précédents pourraient bien entendu s'appliquer. Mais l'hypothèse de complétude simplifie l'algorithme précédent en un algorithme plus efficace. De plus, grâce à cette hypothèse, une politique maximale devient une politique optimale.

Sous ces conditions, de manière similaire à la proposition 4.1, il existe au moins une politique optimale et l'algorithme 4.2 permet de la calculer itérativement.

Proposition 4.3. *Si $\succsim_L$ est complète, transitive et $\succsim_\Phi$ est stable alors pour tout $n > 0$, les ensembles Φ_n^*, Φ_n^+ ne sont pas vides et $\Phi_n^+ \subseteq \Phi_n^*$.*

Si la relation de préférence stricte sur les politiques est stable également, il est possible de construire itérativement toutes les politiques optimales (alg. 4.1).

Proposition 4.4. *Si $\succsim_L$ est complète, transitive et les relations $\succsim_\Phi$ et $\succ_\Phi$ sont stables alors pour tout $n > 0$, Φ_n^* n'est pas vide et $\Phi_n^+ = \Phi_n^*$.*

L'hypothèse de complétude permet de faire le lien entre les propositions 4.1 et 4.3.

Proposition 4.5. *Si $\succsim_L$ est complète, transitive et $\succsim_\Phi$ est stable alors l'égalité suivante est vérifiée :*

$$\forall n > 0, \Phi_n^+ \; = \; \Phi_n^{+M}.$$

Grâce aux propositions précédentes, l'algorithme de recherche arrière généralisé précédent se simplifie :

1: $t \leftarrow N$
2: $\Phi_N^* \leftarrow \{()\}$
3: **repeat**
4: $\quad t \leftarrow t - 1$
5: $\quad$ **for all** $\phi \in \Phi_{t+1}^*$ **do**
6: $\quad\quad$ **for all** $s \in S$ **do**
7: $\quad\quad\quad \Phi_t^*(s) \leftarrow \max(\{(a, \phi) : a \in A\}, \succsim_L)$
8: $\quad\quad$ **end for**
9: $\quad\quad$ ajout dans Φ_t^* des politiques obtenues à partir de $\Phi_t^*(s)$
10: $\quad$ **end for**
11: **until** $t = 0$

Pour chaque politique obtenue à l'étape précédente, les opérations suivantes sont effectuées. Dans chaque état, l'algorithme calcule les meilleures actions à effectuer à l'horizon t (ligne 7), puis construit la ou les meilleures règles de décision pour l'horizon t (ligne 9) en sélectionnant une action parmi la ou les meilleures actions calculées dans chaque état. Ainsi l'algorithme calcule Φ_t^+ à chaque étape. Il repose sur la propriété $\forall n > 0, \Phi_n^+ \subseteq \Phi_n^*$ de la proposition 4.3. Les politiques ainsi construites sont optimales. Quand l'égalité de ces deux ensembles est vérifiée, l'algorithme permet de calculer toutes les politiques optimales.

La différence avec l'algorithme de recherche arrière précédent est la suppression d'une étape de calcul (algo. 4.1, ligne 11). Cette opération n'est plus nécessaire. Et ainsi, pour obtenir une seule politique optimale, il est possible de ne calculer qu'une seule sous-politique optimale à chaque étape. Cette propriété est très intéressante quand on veut déterminer rapidement une politique optimale sans les vouloir toutes.

5. Application à l'utilité bipolaire possibiliste

L'utilité bipolaire possibiliste généralise et unifie les modèles décisionnels que sont les utilités optimistes et pessimistes (Dubois *et al.*, 1998, Dubois *et al.*, 2001). Ces dernières ont été étendues à la décision séquentielle possibiliste par Sabbadin (Sabbadin, 1999). Nous montrons dans cette section que l'utilité bipolaire possibiliste peut également être utilisée pour la décision séquentielle.

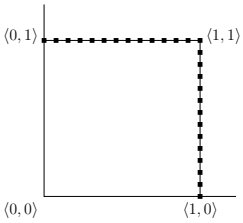

Figure 1. *Echelle d'utilité bipolaire U_V*

On pose $X = \{\langle \lambda, \mu \rangle : \lambda, \mu \in \mathbf{L}, \lambda \vee \mu = 1\}$ (voir fig 1). L'ordre complet $\succeq$ défini sur X est le suivant : $\langle \lambda, \mu \rangle \succeq \langle \lambda', \mu' \rangle \iff \lambda \geq \lambda'$ et $\mu \leq \mu'$. Le plus grand élément de X est donc $\langle 1, 0 \rangle$ et le plus petit $\langle 0, 1 \rangle$. L'opérateur $\max$ sur X se définit ainsi : $\max(\langle \lambda, \mu \rangle, \langle \lambda', \mu' \rangle) = \langle \lambda \vee \lambda', \mu \wedge \mu' \rangle$. L'opérateur $\min$ sur X se définit : $\min(\langle \lambda, \mu \rangle, \langle \lambda', \mu' \rangle) = \langle \lambda \wedge \lambda', \mu \vee \mu' \rangle$. On suppose que $\circ = \min$. Les récompenses sont donc à valeurs dans cet espace X qui est en réalité mono-dimensionnel grâce à la contrainte $\lambda \vee \mu = 1$.

L'opérateur $\vee$ s'étend naturellement sur cet espace :

$$\langle \lambda, \mu \rangle \vee \langle \lambda', \mu' \rangle = \langle \lambda \vee \lambda', \mu \vee \mu' \rangle.$$

Similairement, $\wedge$ peut se définir comme un opérateur $\mathbf{L} \times X \to X$:

$$\alpha \wedge \langle \lambda, \mu \rangle = \langle \alpha \wedge \lambda, \alpha \wedge \mu \rangle.$$

L'utilité bipolaire possibiliste proposée par Giang et Shenoy (Giang *et al.*, 2001) s'écrit alors pour une loterie π :

$$PU(\pi) = \bigvee_{\langle \lambda, \mu \rangle \in X} \pi(\langle \lambda, \mu \rangle) \wedge \langle \lambda, \mu \rangle.$$

La relation de préférence sur les loteries induite par ce critère sera notée $\succsim_{PU}$.

Proposition 5.1. *La relation $\succsim_{PU}$ est transitive, invariante par translation et indépendante.*

Démonstration. La transitivité est évidente. Démontrons l'indépendance. Soit π_1 et π_2 deux loteries telles que $\pi_1 \succsim_{PU} \pi_2$. Prenons π une troisième loterie et λ, μ dans $\mathbf{L}$ tels que $\lambda \vee \mu = 1$. Calculons

$$PU([\lambda/\pi_1, \mu/\pi]) = \bigvee_{x \in X} ((\lambda \wedge \pi_1(x)) \vee (\mu \wedge \pi_2(x))) \wedge x.$$

Par distributivité de $\wedge$ sur $\vee$, on obtient

$$PU([\lambda/\pi_1, \mu/\pi]) = \bigvee_{x \in X} ((\lambda \wedge \pi_1(x)) \wedge x) \vee ((\mu \wedge \pi_2(x)) \wedge x).$$

Par commutativité de $\vee$, on a

$$PU([\lambda/\pi_1, \mu/\pi]) = (\bigvee_{x \in X} (\lambda \wedge \pi_1(x)) \wedge x) \vee (\bigvee_{x \in X} (\mu \wedge \pi_2(x)) \wedge x).$$

Par distributivité de $\wedge$ sur $\vee$, on obtient

$$PU([\lambda/\pi_1, \mu/\pi]) = (\lambda \wedge \bigvee_{x \in X} (\pi_1(x)) \wedge x) \vee (\bigvee_{x \in X} (\mu \wedge \pi_2(x)) \wedge x).$$

Donc

$$PU([\lambda/\pi_1, \mu/\pi]) \succeq (\lambda \wedge PU(\pi_2)) \vee (\bigvee_{x \in X} (\mu \wedge \pi_2(x)) \wedge x).$$

Finalement,

$$PU([\lambda/\pi_1, \mu/\pi]) \succeq PU([\lambda/\pi_2, \mu/\pi]).$$

Passons maintenant à l'invariance par translation. Notons les éléments de X par $\{\langle \lambda_1, \mu_1 \rangle, \ldots, \langle \lambda_k, \mu_k \rangle\}$. Soit π_1 et π_2 deux loteries telles que $\pi_1 \succsim_{PU} \pi_2$. La loterie π_1 s'écrit

$$[\alpha_1/\langle \lambda_1, \mu_1 \rangle, \ldots, \alpha_k/\langle \lambda_k, \mu_k \rangle].$$

La loterie π_2 s'écrit

$$[\beta_1/\langle \lambda_1, \mu_1 \rangle, \ldots, \beta_k/\langle \lambda_k, \mu_k \rangle].$$

Soit $c = \langle c_1, c_2 \rangle$ un élément de X. Rappelons que $\circ$ est l'opérateur $\min$ sur X. Alors

$$PU(\pi_1^{\rightarrow c}) = \langle \bigvee_{i=1}^{k} \alpha_i \wedge c_1 \wedge \lambda_i, \bigvee_{i=1}^{k} \alpha_i \wedge (c_2 \vee \mu_i) \rangle.$$

Par distributivité de $\wedge$ sur $\vee$, on a

$$PU(\pi_1^{\rightarrow c}) = \langle c_1 \wedge \bigvee_{i=1}^{k} \alpha_i \wedge \lambda_i, \bigvee_{i=1}^{k} (\alpha_i \wedge c_2) \vee (\alpha_i \wedge \mu_i) \rangle.$$

Par commutativité,

$$PU(\pi_1^{\rightarrow c}) = \langle c_1 \wedge \bigvee_{i=1}^{k} \alpha_i \wedge \lambda_i, (\bigvee_{i=1}^{k} \alpha_i \wedge c_2) \vee (\bigvee_{i=1}^{k} \alpha_i \wedge \mu_i) \rangle.$$

Par distributivité de $\wedge$ sur $\vee$ et comme $\bigvee_{i=1}^{k} \alpha_i = 1$, on a

$$PU(\pi_1^{\rightarrow c}) = \langle c_1 \wedge \bigvee_{i=1}^{k} \alpha_i \wedge \lambda_i, c_2 \vee (\bigvee_{i=1}^{k} \alpha_i \wedge \mu_i)\rangle.$$

D'où,

$$PU(\pi_1^{\rightarrow c}) \succeq \langle c_1 \wedge \bigvee_{i=1}^{k} \beta_i \wedge \lambda_i, c_2 \vee \bigvee_{i=1}^{k} \beta_i \wedge \mu_i\rangle.$$

Et finalement,

$$PU(\pi_1^{\rightarrow c}) \succeq PU(\pi_2^{\rightarrow c}).$$

□

Par conséquent, la proposition 4.3 nous garantit que les politiques calculées par l'algorithme 4.2 sont optimales pour l'utilité bipolaire possibiliste.

6. Conclusion

Dans le cadre de la décision séquentielle dans l'incertain possibiliste, nous avons proposé des propriétés simples et suffisantes sur la relation de préférence sur les loteries garantissant l'admissibilité de la recherche arrière quand la relation de préférence sur les loteries est transitive, indépendante et invariante par translation. Si, de plus, cette relation est complète, l'algorithme général peut se simplifier.

Dans la pratique, ces résultats peuvent permettre d'identifier rapidement et simplement des structures de préférence compatibles avec l'utilisation de méthodes fondées sur la programmation dynamique, justifiant ainsi l'utilisation des algorithmes généraux (4.1, 4.2). Ces résultats ont été appliqués sur l'utilité bipolaire possibiliste à titre d'exemple.

Dans cet article, nous avons traité les problèmes où l'horizon est fini. Il serait intéressant d'étendre cette étude à l'horizon infini et de rechercher des conditions sur les relations de préférence permettant de garantir l'existence de politiques préférées stationnaires.

Par ailleurs, nous ne nous sommes intéressés ici qu'à des structures de préférence cohérentes dynamiquement. La cohérence dynamique, notion issue des travaux en sciences économiques (Hammond, 1988, Ghirardato, 2002), stipule que si une alternative est préférée à un instant donné, elle le sera vue de tout autre instant. Dans le cadre des PDM, la notion de cohérence dynamique est proche du principe de Bellman selon lequel toute sous-politique d'une politique optimale est optimale. Les développements de la théorie de la décision en économie, ces dernières années, ont montré l'intérêt de recourir à des modèles plus sophistiqués (e.g. utilité fondée sur l'intégrale de Choquet) pour leur aptitude à modéliser des comportements décisionnels plus élaborés. Malheureusement, ces modèles ne sont pas dynamiquement cohérents (McClennen,

1990, Jaffray *et al.*, 2006), ce qui soulève des problèmes computationnels difficiles. En intelligence artificielle, une contrepartie qualitative de l'intégrale de Choquet a été proposée sous la forme de l'intégrale de Sugeno. Il serait intéressant d'exploiter une utilité fondée sur cet intégrale pour la prise de décision séquentielle.

7. Bibliographie

Dubois D., Fargier H., Lang J., Prade H., Sabbadin R., « Qualitative decision theory and multistage decision making : A possibilistic approach », *Proc. of the European Workshop on Fuzzy Decision Analysis for Management, Planning and Optimization (EFDAN'96)*, 1996.

Dubois D., Godo L., Prade H., Zapico A., « Making Decision in a Qualitative Setting : from Decision under Uncertainty to Case-based Decision », *KR*, vol. 6, p. 594-607, 1998.

Dubois D., Prade H., « Possibility Theory as a basis of Qualitative Decision Theory », *IJCAI*, vol. 14, p. 1925-1930, 1995.

Dubois D., Prade H., Sabbadin R., « Decision-theoretic foundations of qualitative possibility theory », *European Journal of Operational Research*, vol. 128, p. 459-478, 2001.

Fishburn P., *Utility theory for decision making*, Wiley, 1970.

Ghirardato P., « Revisiting Savage in a conditional world », *Economic theory*, vol. 20, p. 83-92, 2002.

Giang P., Shenoy P., « A Comparison of Axiomatic Approaches to Qualitative Decision Making Using Possibility Theory », *UAI*, vol. 17, p. 162-170, 2001.

Hammond P., « Consequentialist Foundations for Expected Utility », *Theory and Decision*, vol. 25, p. 25-78, 1988.

Jaffray J., Nielsen T., « Dynamic decision making without expected utility : an operational approach », *European Journal of Operational Research*, vol. 169, p. 226-246, 2006.

McClennen E., *Rationality and dynamic choice : Foundational explorations*, Cambridge university press, 1990.

Sabbadin R., Une approche ordinale de la décision dans l'incertain : axiomatisation, représentation logique et application à la décision séquentielle, PhD thesis, Université Paul Sabatier de Toulouse, 1998.

Sabbadin R., « A possibilistic model for qualitative sequential decision problems under uncertainty in partially observable environments », *UAI*, vol. 15, p. 567-574, 1999.

Sabbadin R., Fargier H., Lang J., « Towards qualitative approaches to multi-stage decision making », *International Journal of Approximate Reasoning*, vol. 19, p. 441-471, 1998.

Sobel M., « Ordinal dynamic programming », *Management science*, vol. 21, p. 967-975, 1975.

von Neumann J., Morgenstern O., *Theory of games and economic behavior*, Princeton university press, 1944.

Weng P., « Processus de décision markoviens et préférences non classiques », *Revue d'intelligence artificielle*, vol. 20, n° 2-3, p. 411-432, 2006.