Processus de décision markoviens et préférences non classiques

Paul Weng

LIP6
Université Paris 6
8 rue du Capitaine Scott
75015 Paris
paul.weng@lip6.fr

RÉSUMÉ. *Le modèle classique des processus de décision markoviens repose implicitement sur une structure de préférence induite par l'existence de coûts scalaires additifs et l'utilisation d'un certain critère d'évaluation des politiques (total, total pondéré, moyenne, ...). Cette structure de préférence s'appuie sur des hypothèses fortes permettant de vérifier les principes de la programmation dynamique. Nous nous intéressons ici à des processus de décision markoviens dont la structure de préférence est non classique et énonçons des propriétés simples et suffisantes sur ces préférences pour l'application de méthodes fondées sur la programmation dynamique. Ainsi ces propriétés délimitent une classe plus large de processus de décision markoviens résolubles par la programmation dynamique.*

ABSTRACT. *The standard model of Markov decision processes implicitly relies on a preference structure induced by the existence of scalar and additive costs and the use of a certain criterion for policy evaluation (total, discounted, average, ...). This preference structure imposes strict hypotheses allowing the use of dynamic programming. We are interested here in Markov decision processes whose preference structure is non-classic and we give simple and sufficient properties on theses preferences for the use of methods based on dynamic programming. So these properties define a larger class of Markov decision processes solvable with dynamic programming techniques.*

MOTS-CLÉS : *processus de décision markovien, préférence non classique, programmation dynamique*

KEYWORDS: *Markov decision process, non-classic preference, dynamic programming*

1. Introduction

Dans le modèle classique des processus de décision markoviens (PDM), les préférences sur les politiques sont induites par l'existence de coûts scalaires et additifs et par le choix d'un critère d'évaluation des politiques généralement linéaire (total, total pondéré, moyenne, ...). Dans ce cadre particulier, les différentes méthodes classiques de résolution des PDMs (recherche arrière, itération de la valeur, itération de la politique, programmation mathématique) permettent de déterminer les politiques optimales au sens du modèle de préférence considéré. Cependant, la classe des structures de préférence relevant du modèle classique ne permet pas de rendre compte de certaines préférences observées dans des situations complexes de décision. En effet, comme le montrent Krantz *et al.* (1971), les préférences ne sont représentables par une fonction coût scalaire et additivement décomposable que si elles satisfont un certain nombre d'hypothèses structurelles restrictives (complétude, associativité, transitivité, préadditivité, propriété archimédienne). Ainsi dans le cadre classique, pour les chemins, on doit supposer que l'on est capable de comparer toutes les paires de chemins dans un état donné, que ces comparaisons sont transitives et que l'intérêt de chaque chemin peut être quantifié.

Comme le soulignent Perny *et al.* (2002), il existe de nombreuses situations réelles dans lesquelles la structure de préférence viole naturellement l'une des hypothèses ci-dessus. On peut mentionner les exemples suivants :

– dans les problèmes où les coûts s'apprécient selon divers points de vue (énergie, distance, sécurité, ...) non nécessairement réductibles à un critère unique, on peut vouloir apprécier l'intérêt d'une action par un vecteur coût. La comparaison d'actions est alors un problème multicritère puisqu'elle revient à comparer des vecteurs de coûts. Dès lors, l'hypothèse de complétude est remise en question dans la mesure où l'existence de conflits entre critères peut laisser certaines paires d'actions incomparables. C'est par exemple le cas lorsque la préférence utilisée est la dominance de Pareto.

– dans les problèmes où les coûts sont difficiles à évaluer, il est souvent préférable de recourir à une échelle qualitative permettant de graduer l'ordre de grandeur des coûts considérés. Par exemple, on peut vouloir qualifier le niveau de risque associé à certaines actions dans certains états sur une échelle à quatre niveaux (Noir : très risqué, Rouge : risqué, Bleu : normal, Vert : faiblement risqué) sans pour autant pouvoir quantifier ces risques. La comparaison d'actions repose donc sur une préférence sur les couleurs ou les ensembles de couleurs non nécessairement représentable par un modèle numérique additif.

Dans toutes ces situations, le modèle classique des PDMs ne convient pas car les préférences à prendre en compte ne sont pas représentables par un critère coût scalaire additif. Il semble donc intéressant d'étudier l'extension à des PDMs exploitant des préférences non classiques.

Le modèle des processus de décision markoviens a été assez peu étudié du point de vue des structures de préférence. A notre connaissance, les travaux les plus généraux dans cette optique sont ceux de Sobel (1975). En identifiant un PDM à un problème

déterministe défini sur les distributions de probabilité sur l'ensemble d'états, il montre que sous certains axiomes il est possible d'appliquer un algorithme de type itération de la politique pour déterminer des politiques optimales. Mais ces résultats sont difficilement appliquables dans la pratique car les politiques sont de forme complexe, définies sur des espaces infinis (distributions de probabilité sur les états, fonctions de l'ensemble des états dans l'ensemble des actions), et de ce fait, n'ont pas d'interprétation évidente.

On peut néanmoins mentionner quelques études utilisant des préférences non classiques dans le cadre des PDMs, comme Yu *et al.* (1998) qui utilisent comme critère d'évaluation la probabilité d'atteindre un certain niveau de récompense, ou encore Cavazos-Cadena *et al.* (2000) qui utilisent un critère d'utilité sensible au risque. Dans ces travaux, l'hypothèse de l'existence de coûts scalaires numériques est conservée.

Par ailleurs, certains travaux, on peut citer notamment (Furukawa, 1965; Viswanathan *et al.*, 1977; White, 1982; Henig, 1983; Novák, 1989; Wakuta, 1992; Wakuta, 1995) utilisent des valuations numériques non scalaires (vecteurs de réels), donnant naissance aux PDMs multicritères. D'autres travaux ont étudié des PDMs exploitant des préférences qualitatives. On peut citer les travaux de Bonet *et al.* (2002). Leur modèle des processus de décision markoviens qualitatifs traite les problèmes pour lesquels l'information sur les données numériques n'est pas assez riche. Ils utilisent alors des ordres de grandeur pour les probabilités et la fonction de coût. Quand l'incertain est modélisé par la théorie des possibilités, (Dubois *et al.*, 1996; Sabbadin, 1998; Sabbadin *et al.*, 1998; Sabbadin, 1999) ont étendu les critères qualitatifs définis axiomatiquement par (Dubois *et al.*, 1995; Dubois *et al.*, 1998; Dubois *et al.*, 2001) à la contrepartie possibiliste du modèle des processus de décision markoviens.

Les processus de décision markoviens font partie d'une classe plus large de problèmes, ceux de décision dynamique dans l'incertain. Ceux-ci constituent un domaine très étudié en économie (Krebs *et al.*, 1979; Hammond, 1988; Machina, 1989; Ghirardato, 2002). Ces auteurs examinent le problème de la cohérence dynamique en relation avec l'utilité espérée. Celle-ci indique que si une alternative est préférée à un instant donné, elle le sera vue de tout autre instant. Dans le cadre des PDMs, la notion de cohérence dynamique est proche du principe de Bellman qui dit que toute sous-politique d'une politique optimale est optimale. Krebs *et al.* (1979) fournissent dans un cadre axiomatique une généralisation de l'utilité espérée. Hammond (1988) justifie l'utilisation de l'utilité espérée dans le contexte de la décision dynamique. Machina (1989) étudie le problème de la cohérence dynamique pour des modèles non fondés sur l'utilité espérée. Ghirardato (2002) fournit une axiomatisation à la manière de Savage soulignant la relation entre cohérence dynamique et utilité espérée. Dans tous ces travaux, il est supposé que les relations de préférence sont complètes et généralement numériquement représentées, hypothèses que nous ne faisons pas dans cet article.

Nous nous proposons d'étudier les possibilités d'extension du cadre classique des PDMs pour la prise en compte d'une classe de préférence plus large dans les problèmes de planification dans l'incertain. Nos travaux se démarquent de ceux évoqués précédemment sur l'un de ces deux points au moins. Nous ne faisons pas nécessaire-

ment l'hypothèse de complétude des préférences. Même lorsque l'hypothèse de complètude est faite, nous ne supposons pas nécessairement l'existence de récompenses additives. Ainsi nous nous attacherons à étudier sous quelles conditions structurelles des extensions de l'algorithme recherche arrière permettent de déterminer tout ou partie des politiques préférées d'un PDM. Cette étude concerne principalement le cadre probabiliste et peut être vue comme une extension dans l'incertain des travaux de P. Perny (2003) concernant la recherche de chemins préférés dans les graphes.

L'article est organisé de la manière suivante. Dans la section 2, nous présentons le modèle général des PDMs que nous étudierons et nous donnons 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 PDM et donnons les conditions suffisantes pour obtenir la propriété de stabilité permettant le fonctionnement de la programmation dynamique. Dans la section 4 qui présente les principaux résultats de cet article, nous proposons des propriétés suffisantes pour une large classe de structures de préférence (cadre des préférences partielles) garantissant l'admissibilité de l'algorithme de recherche arrière généralisé que nous proposons. Puis nous raffinons ces résultats quand la propriété de complétude de la relation de préférence sur les loteries (cadre des préférences complètes) est imposée en plus. Pour les deux classes de préférence, nous indiquons quelques exemples simples. On remarquera que les propositions de cette section sont formulées de telle sorte qu'elles sont indépendantes de la représentation de l'incertain. Ces propositions pourraient alors s'appliquer sous certaines conditions à d'autres types d'incertain, notamment l'incertain possibiliste. Enfin, en annexe, nous détaillons, sur un exemple exploitant des préférences qualitatives, le déroulement de algorithme général 4.1.2.

2. Cadre général de l'étude

2.1. *Processus de décision markoviens généralisés*

Le modèle général des processus de décision markoviens (PDMG) que nous étudierons est défini par la donnée du quadruplet suivant :

– S l'ensemble des états,

– A l'ensemble des actions,

– T: $S \times A \rightarrow \mathbf{L}(S)$ la fonction de transition où $\mathbf{L}(S)$ est l'ensemble des distributions de probabilité sur S,

– C: $S \times A \times S \rightarrow (X, \circ, \succsim)$ la fonction générale de coûts où X est l'ensemble de valuation "abstraite" des coûts.

Nous supposons dans notre cadre de travail que l'ensemble des états S et l'ensemble des actions A sont finis. L'ensemble des coûts X est muni d'un opérateur interne $\circ$ et d'une relation de préférence $\succsim$. De plus, pour simplifier les notations, cet opérateur $\circ$ est supposé associatif et simplifiable à gauche (i.e. $\forall x, y, z \in X, x \circ y = x \circ z \Rightarrow y = z$).

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 $(X, \circ, \succsim) = (\mathbf{R}, +, \geq)$, on a alors $z \bullet x = z - x$. D'après l'hypothèse de simplifiabilité à gauche, quand cet ensemble n'est pas vide, c'est un singleton.

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

$$(s, a_1, s_1, a_2, s_2, \cdots) \text{ où } \forall i \in \mathbf{N}, (a_i, s_i) \in A \times S.$$

On note les ensembles d'historiques débutant de l'état s par

$$\forall n > 0, \Gamma_n^s = \{(s, a_1, s_1, a_2, s_2, \cdots, a_n, s_n) \mid \forall i = 1, \ldots, n, (a_i, s_i) \in A \times S\}.$$

La valeur ou le coût abstrait d'un historique $\gamma \in \Gamma_n^{s_0}$ débutant à l'état s_0 avec $\gamma = (s_0, a_1, s_1, a_2, s_2, \cdots, a_n, s_n)$ vaut $x = x_1 \circ \cdots \circ x_n \in X$ où $\forall i = 1 \cdots n, x_i = C(s_{i-1}, a_i, s_i)$. La structure $(X, \circ, \succsim)$ est choisie de telle sorte qu'elle représente les préférences sur les historiques. La relation $\succsim$ de l'ensemble X correspond donc à la relation de préférence sur les historiques. Nous noterons indifféremment les deux relations $\succsim$. Posons $\Gamma = \bigcup_n \bigcup_{s \in S} \Gamma_n^s$. Cet ensemble contient tous les historiques potentiellement réalisables pour chaque horizon.

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 $\phi_n = (\delta_1, \cdots, \delta_n)$ où chaque $\delta_i \in \Delta$. 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 $(a, \Phi) = \{(a, \phi) \mid \phi \in \Phi\}$. Par convention, $(a, \emptyset) = \{(a)\}$.

Remarquons qu'une règle de décision δ pour un état s définit une loterie sur l'ensemble X. Cette loterie est égale à $T(s, \delta(s))$. 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 des coûts X induite par la politique ϕ_n à l'état s. Elle associe à tout $x \in X$ la probabilité :

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

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

Dans le cadre classique avec la structure $(\mathbf{R}, +, \geq)$, cette probabilité s'écrit ainsi :

$$L_s^{\phi_n}(x) = \sum_{s' \in S} T(s, \delta(s))(s') L_{s'}^{\phi_{n-1}}(x - C(s, \delta(s), s'))$$

Autrement dit, la probabilité que la politique ϕ_n à l'horizon n génère le coût x est la moyenne pondérée des probabilités aux états s' que la sous-politique ϕ_{n-1} en ces états génère le coût x réduit du coût $C(s, \delta(s), s')$ imputé à l'étape n.

Il est donc possible d'étudier ce modèle selon les propriétés de cet ensemble X. On constate que si l'on prend $(X, \circ, \succsim) = (\mathbf{R}, +, \geq)$, on retrouve le cadre classique des PDMs. Si l'on prend $(X, \circ, \succsim) = (\mathbf{R}^p, +, \geq_D)$ pour $p > 0$, on obtient alors le modèle des PDMs multicritères avec la relation de dominance classique de Pareto $\geq_D$. Avec $X = S \times A \times S$, le PDMG correspond au modèle de Sobel (1975).

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 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. Dans cette dernière définition, l'opération de maximisation est définie une fois sur tout un ensemble contrairement à la définition précédente où la maximisation est définie sur plusieurs petits ensembles. On peut donc soupçonner un coût de calcul plus important pour cette dernière définition. De plus, 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 :

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

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

Définition 2.1. *La relation de préférence $\succsim$ sur l'ensemble $(X, \circ)$ est dite* préadditive *si et seulement si pour tout $\gamma, \gamma' \in X$, pour tout $x \in X$,*

$$\gamma \succsim \gamma' \Leftrightarrow x \circ \gamma \succsim x \circ \gamma'.$$

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 probabilistes sur X. Cette propriété peut être considérée comme la version probabiliste de la préadditivité. En effet, une forme de préadditivité faible est obtenue en considérant les loteries dégénérées.

Définition 2.2. *Une relation de préférence $\succsim_L$ sur les loteries définies sur $(X, \circ)$ est* invariante par translation *si et seulement si pour tout $L_1, L_2 \in \mathbf{L}(X)$, pour tout $c \in X$,*

$$(L_1 \succsim_L L_2 \Rightarrow L_1^{\rightarrow c} \succsim_L L_2^{\rightarrow c})$$

où $\forall i = 1, 2, \forall x \in X, L_i^{\rightarrow c}(c \circ x) = L_i(x)$.

Nous introduisons maintenant la propriété d'indépendance. Elle correspond en fait à une version affaiblie de la propriété d'indépendance de l'axiomatique de von Neumann *et al.* (1944) formulée par 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 probabilités) à deux loteries ne peut inverser le sens de préférence.

Définition 2.3. *Une relation de préférence $\succsim_L$ sur les loteries vérifie la propriété d'*indépendance *si et seulement si pour tout $L_1, L_2, L_3 \in \mathbf{L}(X)$, pour tout $\lambda \in]0, 1[$,*

$$L_1 \succsim_L L_2 \Rightarrow \lambda L_1 + (1 - \lambda) L_3 \succsim_L \lambda L_2 + (1 - \lambda) L_3).$$

Nous définissons 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 2.4. *Une relation de préférence $\succsim_\Phi$ sur les politiques sera dite* stable *si et seulement si pour tout $\phi, \phi' \in \Phi$, pour tout $\delta \in \Delta$,*

$$(\phi \succsim_\Phi \phi' \Rightarrow (\delta, \phi) \succsim_\Phi (\delta, \phi')).$$

Considérons pour $\delta \in \Delta$, l'opérateur $H_\delta \colon \Phi \to \Phi$ 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 dans un PDMG

Dans le modèle des PDMs ou des PDMGs, il est possible de distinguer trois niveaux de relations de préférence. Une première relation $\succsim$ est définie sur les historiques ou de manière équivalente sur l'ensemble des coûts X. Comme une politique pour un horizon fixé et un état initial donné définit une loterie sur l'ensemble X, comparer deux politiques à un horizon donné et dans un certain état initial équivaut à comparer leurs loteries respectives. C'est pourquoi à partir de la première relation de préférence, il est nécessaire de définir une relation de préférence $\succsim_L$ sur les loteries. 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

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

Voici deux lemmes qui nous serviront ultérieurement.

Lemme 3.1. *Si $\succsim_L$ est transitive alors $\succsim_\Phi$ est transitive. De plus, si $\succsim_\Phi$ est stable alors la relation $\sim_\Phi$ est stable également.*

Le lemme suivant 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.

Lemme 3.2. *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 $\forall i = 1, \ldots, n, L_i \succsim_L L'_i$, on a*

$$\forall i = 1, \ldots, n, \lambda_i \in [0,1], \text{ tels que } \sum_{i=1}^n \lambda_i = 1, \sum_{i=1}^n \lambda_i L_i \succsim_L \sum_{i=1}^n \lambda_i L'_i.$$

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 $\forall \lambda \in [0,1], \lambda L_1 + (1-\lambda)L_2 \succsim_L \lambda L'_1 + (1-\lambda)L_2$. Puis en appliquant la propriété d'indépendance sur la seconde relation et L'_1, on a $\forall \lambda \in [0,1], \lambda L'_1 + (1-\lambda)L_2 \succsim_L \lambda L'_1 + (1-\lambda)L'_2$. Enfin par transitivité, on obtient bien : $\forall \lambda \in [0,1], \lambda L_1 + (1-\lambda)L_2 \succsim_L \lambda L'_1 + (1-\lambda)L'_2$.

Supposons 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 $\forall i = 1, \ldots, n+1, L_i \succsim_L L'_i$. Soit une séquence $(\lambda_i)_{i=1..n+1} \in [0,1]$ telle que $\sum_{i=1..n+1} \lambda_i = 1$.

Cas 1 : $\lambda_{n+1} = 1$: La propriété est démontrée.

Cas 2 : $\lambda_{n+1} \neq 1$: Posons $L = \sum_{i=1..n} \lambda_i/(1-\lambda_{n+1})L_i$ et $L' = \sum_{i=1..n} \lambda_i/(1-\lambda_{n+1})L'_i$. 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 $\lambda = \lambda_{n+1}$, on obtient :

$$\lambda_{n+1}L_{n+1} + (1-\lambda_{n+1})L \succsim_L \lambda_{n+1}L'_{n+1} + (1-\lambda_{n+1})L'.$$

En développant L et L', on obtient bien : $\sum_{i=1..n} \lambda_i L_i \succsim_L \sum_{i=1..n} \lambda_i L'_i$. □

La proposition suivante donne 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 invariante par translation, transitive 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

$$\forall x \in X, L_s^{(\delta,\phi)}(x) = \sum_{s' \in S} T(s, \delta(s))(s') L^{\phi}_{s'}(x \bullet C(s, \delta(s), s')).$$

De même, pour (δ, ϕ'), on obtient

$$\forall x \in X, L_s^{(\delta,\phi')}(x) = \sum_{s' \in S} T(s, \delta(s))(s') L^{\phi'}_{s'}(x \bullet C(s, \delta(s), s')).$$

En posant $\forall s' \in S, \forall x \in X, L_{s'}(x) = L^{\phi}_{s'}(x \bullet C(s, \delta(s), s'))$ et $L'_{s'}(x) = L^{\phi'}_{s'}(x \bullet C(s, \delta(s), s'))$, on peut réécrire les loteries $L_s^{(\delta,\phi)} = \sum_{s' \in S} T(s, \delta(s))(s') L_{s'}$ et $L_s^{(\delta,\phi')} = \sum_{s' \in S} T(s, \delta(s))(s') L'_{s'}$. D'après l'hypothèse de simplifiabilité à gauche,

les ensembles $x \bullet C(s, \delta(s), s')$ sont des singletons ou sont vides. Pour les x tels que $x \bullet C(s, \delta(s), s')$ est vide, $L_{s'}(x) = L^{\phi}_{s'}(\emptyset) = 0$. Pour les x tels que $x \bullet C(s, \delta(s), s')$ est un singleton, il existe un y tel que $x \bullet C(s, \delta(s), s') = y$, autrement dit, $x = C(s, \delta(s), s') \circ y$. Alors, par définition, $L_{s'}(x) = L_{s'}(C(s, \delta(s), s') \circ y) = L^{\phi}_{s'}(y)$. Plus simplement, on peut donc écrire, $\forall s' \in S, \forall x \in X, L^{\phi}_{s'}(x) = L_{s'}(C(s, \delta(s), s') \circ x)$ et $L^{\phi'}_{s'}(x) = L'_{s'}(C(s, \delta(s), s') \circ x)$. Donc en vertu de l'hypothèse d'invariance par translation, $\forall s' \in S, L_{s'} \succsim_L L'_{s'}$. D'après le lemme précédent 3.2, $\sum_{s' \in S} T(s, \delta(s))(s') L_{s'} \succsim_L \sum_{s' \in S} T(s, \delta(s))(s') L'_{s'}$. 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. Etude de deux structures de préférence

Dans cette section, nous énonçons en premier lieu nos résultats dans un cadre général (cadre des préférences partielles) garantissant que des politiques préférées existent et peuvent être construites itérativement par recherche arrière (algo. 4.1.2). Ensuite nous listons quelques exemples entrant dans ce cadre. Puis nous affinons les résultats obtenus dans le cas particulier des préférences complètes. Nous établissons le lien avec les résultats précédents et fournissons une spécification (algo. 4.2.2) plus efficace de l'algorithme général précédent. Enfin, quelques exemples sont également présentés pour cette classe de préférence.

4.1. *Cadre des préférences partielles*

4.1.1. *Résultats*

Le cadre des préférences partielles se caractérise par la donnée d'une relation de préférence transitive sur les loteries et d'une relation de préférence stable sur les politiques. Il inclut notamment le modèle des PDMs multicritères. Sous ces conditions, nous démontrons qu'il existe au moins une politique maximale et que l'algorithme 4.1.2 permet de la calculer itérativement.

Les trois lemmes suivants nous serviront ultérieurement pour les démonstrations.

Lemme 4.1. *Quand $\succ_\Phi$ est stable, on a $\forall n > 0, \Phi^{+M}_n = M(\Phi^+_n)$.*

Démonstration. Le résultat se démontre par récurrence sur n. Pour $n = 1$, c'est vrai par définition.

Supposons que l'égalité est vraie à l'étape n. Par définition, on a

$$\Phi^+_{n+1} = \bigcup_{\phi_n \in \Phi^+_n} M(\{(\delta, \phi_n) \mid \delta \in \Delta\}).$$

Donc, on obtient

$$M(\Phi^+_{n+1}) = M(\bigcup_{\phi_n \in \Phi^+_n} M(\{(\delta, \phi_n) \mid \delta \in \Delta\})).$$

Par conséquent,

$$M(\Phi^+_{n+1}) = M(\bigcup_{\phi_n \in \Phi^+_n} \{(\delta, \phi_n) \mid \delta \in \Delta\}).$$

De plus, on a

$$M(\Phi^+_{n+1}) = M(\bigcup_{\phi_n \in M(\Phi^+_n)} \{(\delta, \phi_n) \mid \delta \in \Delta\})$$

car les éléments dominés de Φ^+_n par addition d'une règle δ seront dominés en vertu de la stabilité de $\succ_\Phi$.

D'après l'hypothèse de récurrence,

$$M(\Phi^+_{n+1}) = M(\bigcup_{\phi_n \in \Phi^{+M}_n} \{(\delta, \phi_n) \mid \delta \in \Delta\}).$$

D'où $M(\Phi^+_{n+1}) = \Phi^{+M}_{n+1}$. □

Lemme 4.2. *Soit $(X, \succsim)$ un ensemble partiellement ordonné. Soient A, B deux sous-ensembles de X. Si $M(B) \subseteq A \subseteq B$ alors $M(A) = M(B)$.*

Démonstration. $M(B) \subseteq M(A)$ **:** Soit $b \in M(B)$. Donc $\forall a \in B, \neg(a \succ b)$. Donc en particulier, $\forall a \in A, \neg(a \succ b)$. D'où $b \in M(A)$.

$M(A) \subseteq M(B)$ **:** Soit $a \in M(A)$. Supposons qu'il existe un $c \in B, c \succ a$. Il existe alors un $b \in M(B), b \succsim c$. Donc $b \succ a$ par transitivité. Or $b \in A$. Il y a alors contradiction avec $a \in M(A)$. Par conséquent, $a \in M(B)$.

□

Lemme 4.3. *Si $\succsim_\Phi$ est stable et transitive alors pour tout $n > 0, \forall \phi^*_n \in \Phi^*_n$, il existe $\phi^+_n \in \Phi^{+M}_n$ telle que $\phi^{+M}_n \sim_\Phi \phi^*_n$.*

Démonstration. Démontrons par récurrence sur n.

Pour $n = 1$, c'est vrai par définition.

Soit $n \geq 1$. On suppose que la propriété suivante est vérifiée :

$$\forall \phi^*_n \in \Phi^*_n, \exists \phi^+_n \in \Phi^{+M}_n, \phi^{+M}_n \sim_\Phi \phi^*_n.$$

Montrons que cette propriété est vraie pour $n + 1$ également.

Soit $\phi^*_{n+1} = (\delta^*, \phi^*_n) \in \Phi^*_{n+1}$. Par hypothèse de récurrence, il existe $\phi^+_n \in \Phi^{+M}_n$ telle que $\phi^+_n \sim_\Phi \phi^*_n$. D'après le lemme 3.1, $(\delta^*, \phi^+_n) \sim_\Phi (\delta^*, \phi^*_n)$. Donc $(\delta^*, \phi^+_n) \in \Phi^*_n$ et $(\delta^*, \phi^+_n) \in \Phi^{+M}_n$. □

Notre résultat pour le cadre des préférences partielles s'énonce ainsi : 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_n^{+M} ne sont pas vides et $\Phi_n^{+M} \subseteq \Phi_n^*$.*

Démonstration. D'après le lemme 3.1, $\succsim_\Phi$ est transitive.

La démonstration se fait par récurrence sur n.

Pour $n = 1$, pour chaque état, on peut sélectionner une action maximale. On peut définir par conséquent une règle de décision maximale. On a alors $\Phi_1^{+M} = \Phi_1^*$ qui sont non vides.

Soit $n \geq 1$. On suppose que les ensembles Φ_n^*, Φ_n^{+M} sont non vides et $\Phi_n^{+M} \subseteq \Phi_n^*$. Démontrons ce résultat pour $n + 1$.

Par construction, Φ_{n+1}^{+M} est non vide également. Soit $\phi_{n+1}^+ = (\delta^+, \phi_n^+) \in \Phi_{n+1}^{+M}$ avec $\delta^+ \in \Delta$ et $\phi_n^+ \in \Phi_n^{+M} \subseteq \Phi_n^*$. Montrons qu'elle est dans Φ_{n+1}^*.

Par l'absurde, supposons qu'il existe $\phi_{n+1}^- = (\delta^-, \phi_n^-) \in \Phi_{n+1}$ telle que $\phi_{n+1}^- \succ_\Phi \phi_{n+1}^+$.

Il existe $\phi_n^* \in \Phi_n^*$ telle que $\phi_n^* \succsim_\Phi \phi_n^-$. D'après le lemme 4.3, il existe $\phi_n' \in \Phi_n^{+M}$ telle que $\phi_n' \sim_\Phi \phi_n^*$. D'où $\phi_n' \succsim_\Phi \phi_n^-$. Par stabilité, $(\delta^-, \phi_n') \succsim_\Phi (\delta^-, \phi_n^-)$. Par transitivité, $(\delta^-, \phi_n') \succ_\Phi \phi_{n+1}^+$.

Il existe ϕ_{n+1}'' telle que $\phi_{n+1}'' \in \Phi_{n+1}^{+M}$ et $\phi_{n+1}'' \succsim_\Phi (\delta^-, \phi_n')$. Par transitivité, $\phi_{n+1}'' \succ_\Phi \phi_{n+1}^+$. Il y a donc contradiction.

Finalement, $\forall \phi_{n+1} \in \Phi_{n+1}, \neg(\phi_{n+1} \succ_\Phi \phi_{n+1}^+)$ et $\phi_{n+1}^+ \in \Phi_{n+1}^*$. □

Si 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 peuvent être calculées de manière itérative.

Proposition 4.2. *Si $\succ_\Phi$ est stable alors $\Phi_n^* \subseteq \Phi_n^+$.*

Démonstration. La démonstration se fait par récurrence sur n.

Pour $n = 1$, la relation est vérifiée par définition.

Soit $n \geq 1$. On suppose que $\Phi_n^* \subseteq \Phi_n^+$. Démontrons cette inclusion pour $n + 1$.

Soit $\phi_{n+1}^* = (\delta^*, \phi_n^*) \in \Phi_{n+1}^*$. Par l'absurde, supposons qu'il existe $\phi_n \in \Phi_n$ telle que $\phi_n \succ_\Phi \phi_n^*$. Par stabilité, on a $(\delta^*, \phi_n) \succ_\Phi (\delta^*, \phi_n^*)$. Il y a donc contradiction.

D'où, $\forall \phi_n \in \Phi_n, \neg(\phi_n \succ_\Phi \phi_n^*)$. Donc $\phi_n^* \in \Phi_n^* \subseteq \Phi_n^+$. De plus, par définition de ϕ_{n+1}^*, on a $\forall \delta \in \Delta, \neg((\delta, \phi_n^*) \succ_\Phi (\delta^*, \phi_n^*)$. Par conséquent, $\phi_{n+1}^* \in \Phi_{n+1}^+$. □

Par conséquent, ces deux dernières propositions (4.1 et 4.2) énoncent les conditions suffisantes sur les relations de préférence pour que la méthode de recherche arrière généralisée (algo. 4.1.2) permette de déterminer toutes les politiques maximales.

Corollaire 4.4. *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 $\Phi_n^{+M} = \Phi_n^*$.*

Démonstration. D'après le lemme 4.1, on a $\Phi_n^{+M} = M(\Phi_n^+)$ car $\succ_\Phi$ est stable. D'après la proposition 4.2, $\Phi_n^* \subseteq \Phi_n^+$ grâce à la stabilité de $\succ_\Phi$. D'où d'après le lemme 4.2, en posant $A = \Phi_n^*$ et $B = \Phi_n^+$, l'égalité est démontrée. □

4.1.2. *Algorithme de recherche arrière généralisé (version 1)*

L'algorithme de recherche arrière généralisé (version 1) s'écrit :

1: $t \leftarrow N$
2: $\Phi_N^{+M} \leftarrow \{()\}$
3: **repeat**
4: $\quad t \leftarrow t - 1$
5: $\quad$ **for all** $\phi \in \Phi_{t+1}^{+M}$ **do**
6: $\quad\quad$ **for all** $s \in S$ **do**
7: $\quad\quad\quad \Phi_t^{+M}(s) \leftarrow \Phi_t^{+M}(s) \bigcup M(\{(a, \phi) : a \in A\})$
8: $\quad\quad$ **end for**
9: $\quad\quad$ ajout dans Φ_t^{+M} des politiques obtenues à partir de $\Phi_t^{+M}(s)$
10: $\quad$ **end for**
11: $\quad \Phi_t^{+M} \leftarrow M(\Phi_t^{+M})$
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 cet horizon (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}. Il repose sur la propriété suivante : $\forall t > 0, \Phi_t^* = \Phi_t^{+M}$. Quand seule la propriété $\forall t > 0, \Phi_t^{+M} \subseteq \Phi_t^*$ est vérifiée, une politique maximale peut encore être calculée itérativement. Mais il n'est plus possible de les obtenir toutes.

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 $t < N$. On remarquera que c'est le cas pour les PDMs multicritères.

On constate qu'on a changé d'espace de travail par rapport au modèle standard des PDMs qui utilise l'espace de valuation (les réels) pour évaluer les actions. L'al-

gorithme 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 coûts généralisés que peut engendrer une politique donnée et les probabilités associées à ces coûts. 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 (espérance de la somme des coûts, par exemple), l'algorithme proposé se simplifie naturellement (lignes 7 et 11).

4.1.3. *Exemples*

Les conditions de la proposition 4.1 sont générales. De nombreuses structures de préférence les vérifient. A titre d'exemple, nous en citons trois.

Considérons un PDM classique dont les coûts sont définis sur $(\mathbf{R}, +, <)$. La relation de préférence sur les loteries est définie par la relation de dominance stochastique de premier ordre. En utilisant les mêmes arguments que dans la preuve de la proposition 4.3, on peut montrer que la relation de dominance stochastique de premier ordre est transitive, invariante par translation et indépendante. La structure de préférence ainsi définie vérifie donc les conditions de la proposition 4.1. En fait, pour cette relation, la proposition 4.2 s'applique également.

Les PDMs dont le coût est mesuré par un vecteur de réels exploitant le critère total, total pondéré ou moyenne forment un autre exemple de structures de préférence vérifiant les conditions des propositions 4.1 et 4.2, la relation de préférence sur les vecteurs étant simplement la relation de dominance de Pareto. Par conséquent, comme pour l'exemple précédent, l'instanciation de l'algorithme 4.1.2 sur cette structure permet de calculer toutes les politiques non dominées.

Enfin, nous présentons un exemple un peu plus détaillé que nous reprendrons en annexe pour montrer le déroulement de l'algorithme 4.1.2. Considérons le problème de navigation d'un robot autonome dans un environnement hostile. L'environnement est modélisé par une grille. Les états sont alors la position du robot dans la grille. Les actions sont les déplacements possibles du robot. A chaque position de la grille est affecté un niveau de risque. Celui-ci est difficilement quantifiable et on le mesure sur une échelle qualitative : Noir (très risqué), Rouge (risqué), Bleu (normal), Vert (risque faible). Les coûts sont donc qualitatifs. Naturellement, le coût Vert est préféré à Bleu qui est préféré au coût Rouge qui est, lui même, préféré à Noir. L'ensemble des coûts X contient ces quatre couleurs et toutes les séquences composées de ces quatres couleurs. L'opérateur $\circ$ sur X est simplement la concaténation. La relation de préférence $\trianglerighteq$ sur cet ensemble de coûts est définie par la relation de Bossong *et al.* (1997) : une séquence de couleurs $\gamma = (x_1, \cdots, x_k)$ sera préférée à une autre séquence de couleurs $\gamma' = (x'_1, \cdots, x'_l)$ si et seulement si $k \leq l$ et il existe une injection i de $\{x_1, \cdots, x_k\}$ dans $\{x'_1, \cdots, x'_l\}$ telle que $\forall x \in \{x_1, \cdots, x_k\}$, la couleur x est préférée à la couleur

$i(x)$. Cette relation est préadditive (Spanjaard, 2003). La partie stricte de cette relation sera notée $\rhd$.

La relation de préférence sur les loteries $\succsim_{DS}$ est définie par $L \succsim_{DS} L'$ si et seulement si $\forall k \in X, \sum_{x \in X, \neg(x \lhd k)} L(x) \geq \sum_{x \in X, \neg(x \lhd k)} L'(x)$. Littéralement, cela signifie que la probabilité pour la loterie L de l'évènement "obtenir des conséquences qui ne soient pas moins préférées qu'un niveau donné" est plus grande que la probabilité de ce même évènement pour la loterie L'. Cette relation peut être vue comme une généralisation de la relation de dominance stochastique de premier ordre au cas d'un ensemble de conséquences partiellement ordonné. Cette relation est partielle et induit une relation de préférence sur les politiques stable sous l'hypothèse de préadditivité.

Proposition 4.3. *Les relations $\succsim_{DS}$ et $\succ_{DS}$ sont transitives, indépendantes et invariantes par translation.*

Démonstration. La transitivité et l'indépendance de $\succsim_{DS}$ sont évidentes. Montrons l'invariance par translation.

Soit un couple de loteries L_1, L_2 tel que $L_1 \succsim_{DS} L_2$. Par définition, cela s'écrit $\forall k \in X, \sum_{\neg(x \lhd k)} L_1(x) \geq \sum_{\neg(x \lhd k)} L_2(x)$.

Soit $c \in X$. Définissons $\forall x \in X, L'_1(c \circ x) = L_1(x)$ et $\forall x \in X, L'_2(c \circ x) = L_2(x)$. On peut écrire $\forall k \in X, \sum_{\neg(x \lhd k)} L'_1(c \circ x) \geq \sum_{\neg(x \lhd k)} L'_2(c \circ x)$. Par préadditivité, $\forall k \in X, \sum_{\neg(c \circ x \lhd c \circ k)} L'_1(c \circ x) \geq \sum_{\neg(c \circ x \lhd c \circ k)} L'_2(c \circ x)$. Donc $\forall k \in X, \sum_{\neg(y \lhd c \circ k)} L'_1(y) \geq \sum_{\neg(y \lhd c \circ k)} L'_2(y)$. Enfin, $\forall k' = c \circ k \in c \circ X \subseteq X, \sum_{\neg(y \lhd k')} L'_1(y) \geq \sum_{\neg(y \lhd k')} L'_2(y)$.

On procède de même pour la relation stricte $\succ_{DS}$. □

D'après les propositions 3.1 et 4.1, on sait donc que dans un PDM muni des deux relations de préférence $\trianglerighteq$ et $\succsim_{DS}$, il est possible d'utiliser l'algorithme de recherche arrière généralisé 4.1.2 pour déterminer une politique maximale.

4.2. *Le cadre des préférences complètes*

4.2.1. *Résultats*

Le cadre des préférences complètes se caractérise par la donnée d'une relation de préférence complète et transitive sur les loteries et d'une relation de préférence stable sur les politiques. Par rapport à la structure de préférence étudiée précédemment, on ajoute l'hypothèse de complétude de la relation de préférence sur les loteries. Les résultats précédents pourraient bien entendu s'appliquer. Mais, l'hypothèse de complétude permet de simplifier l'algorithme précédent et d'obtenir un algorithme plus efficace. De plus, grâce à cette hypothèse, une politique maximale est une politique optimale.

Sous ces conditions, de manière similaire à la proposition 4.1, nous démontrons qu'il existe au moins une politique optimale et que l'algorithme 4.2.2 permet de la calculer itérativement.

Proposition 4.4. *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 $\Phi^+_n \subseteq \Phi^*_n$.*

Démonstration. D'après le lemme 3.1, la relation $\succsim_\Phi$ est transitive.

La démonstration se fait par récurrence sur n.

Pour $n = 1$, pour chaque état, on peut sélectionner une meilleure action car $\succsim_L$ est complète. On peut définir par conséquent une meilleure règle de décision. On a alors $\Phi^+_1 = \Phi^*_1$ qui sont non vides.

Soit $n \geq 1$. On suppose que les ensembles Φ^*_n, Φ^+_n sont non vides et que $\Phi^+_n \subseteq \Phi^*_n$. Démontrons-le pour $n + 1$.

Par construction, Φ^+_{n+1} est non vide également. Soit $\phi^+_{n+1} = (\delta^+, \phi^*_n) \in \Phi^+_{n+1}$ avec $\delta^+ \in \Delta$ et $\phi^*_n \in \Phi^+_n \subseteq \Phi^*_n$. Montrons qu'elle est dans Φ^*_{n+1}.

Par hypothèse, $\forall \phi_n \in \Phi_n, \phi^*_n \succsim_\Phi \phi_n$. Par stabilité, $\forall \phi_n \in \Phi_n, \forall \delta \in \Delta, (\delta, \phi^*_n) \succsim_\Phi (\delta, \phi_n)$. Or, comme $\succsim_L$ est complète, par définition de $\phi^+_{n+1}, \forall \delta \in \Delta, (\delta^+, \phi^*_n) \succsim_\Phi (\delta, \phi^*_n)$. Donc par transitivité, $\forall \phi_n \in \Phi_n, \forall \delta \in \Delta, (\delta^+, \phi^*_n) \succsim_\Phi (\delta, \phi_n)$. Par conséquent, $\phi^+_{n+1} \in \Phi^*_{n+1}$ et cet ensemble est non vide. □

Le corollaire suivant montre que dans le cadre des préférences complètes, quand la relation de préférence stricte sur les politiques est stable également, il est possible de construire itérativement toutes les politiques optimales (alg. 4.2.2). En effet, dans le modèle des PDMs standard, la relation de préférence stricte sur les politiques est stable et l'algorithme 4.2.2 qui devient la méthode usuelle de recherche arrière, peut en effet calculer toutes les politiques optimales.

Corollaire 4.5. *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 4.4.

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

$$\forall n > 0, \Phi^+_n = \Phi^{+M}_n.$$

Démonstration. La démonstration se fait par récurrence sur n.

L'égalité est vraie pour $n = 1$.

Soit $n \geq 1$. Supposons l'égalité vérifiée. Démontrons-la pour $n + 1$.

Par définition,

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

D'après la proposition 4.4, $\Phi_{n+1}^{+} \subset \Phi_{n+1}^{*}$. Par conséquent, $M(\Phi_{n+1}^{+}, \succsim_\Phi) = \Phi_{n+1}^{+}$. Finalement, $\Phi_{n+1}^{+M} = \Phi_{n+1}^{+}$. □

4.2.2. *Algorithme de recherche arrière généralisé (version 2)*

L'algorithme de recherche arrière généralisé (version 2) s'écrit :

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: $\quad\quad\quad \Phi_t^*(s) \leftarrow \Phi_t^*(s) \bigcup \max\{(a, \phi) : a \in A\}$
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 suivante : $\forall t > 0, \Phi_t^* = \Phi_t^+$.

Quand cette dernière propriété est relachée et que seule la relation $\forall n > 0, \Phi_n^+ \subseteq \Phi_n^*$ est vraie, il est encore possible de calculer itérativement une politique optimale. Mais il n'est plus possible de les obtenir toutes. C'est notamment le cas dans la contrepartie possibiliste des PDMs développée par (Dubois *et al.*, 1996; Sabbadin, 1998; Sabbadin *et al.*, 1998; Sabbadin, 1999) pour lequel, seule la proposition 4.4 est valide, la proposition 4.2 ne s'appliquant pas.

La différence avec l'algorithme de recherche arrière précédent est la suppression d'une étape de calcul (algo. 4.1.2, 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 calculer rapidement une politique optimale sans les vouloir toutes.

4.2.3. *Exemples*

Les PDMs classiques sont un exemple de la classe de préférence concernée par la proposition 4.4. Dans un cadre possibiliste, la contrepartie possibiliste des PDMs étudiée par (Dubois *et al.*, 1996; Sabbadin, 1998; Sabbadin *et al.*, 1998; Sabbadin, 1999) en est un autre exemple.

Le modèle des PDMs qualitatifs de Bonet *et al.* (2002) pourrait également être vu comme un exemple du cadre complet. Les probabilités et les coûts sont qualitatifs et sont définis sur l'ensemble des "réels étendus" (Wilson, 1995). Les propriétés dans ces PDMs découlent de celles des PDMs classiques.

Les PDMs utilisant le critère maximin forment un autre exemple de cette classe de préférence complète. Le critère maximin consiste à valuer une loterie par sa plus mauvaise conséquence. Ce critère pessimiste permet le calcul de politiques préférées de manière itérative.

Enfin, les PDMs multicritères où la relation de préférence sur les vecteurs est un ordre lexicographique et où la relation de préférence sur les loteries est représentée par le critère total, total pondéré ou moyenne définissent également une structure de préférence vérifiant la proposition 4.4.

5. Conclusion

Nous avons proposé des propriétés simples et suffisantes sur la relation de préférence sur les loteries garantissant l'admissibilité de la recherche arrière sur deux structures de préférence. La première est caractérisée par la transitivité, l'indépendance et l'invariance par translation de la relation de préférence sur les loteries. La seconde structure est obtenue en imposant en plus la propriété de complétude. Cette dernière condition permet d'obtenir un algorithme moins calculatoire. Pour ces deux classes de préférence, nous avons proposé un algorithme de résolution.

Dans la pratique, ces résultats pourraient 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 (4.1.2, 4.2.2).

Enfin remarquons que nous avons énoncé nos résultats dans un cadre probabiliste. Ils pourraient probablement être transposés au cadre possibiliste (et à d'autres types d'incertain). Comme nous l'avons déjà souligné, le cadre classique (section 4.2.1) contiendrait alors la version possibiliste des PDMs développée par Sabbadin (1998). Le cadre de la section 4.1.1 permettrait une généralisation à des structures de préférence partielles de ce modèle.

Remerciements

Je suis reconnaissant à Patrice Perny, Jean-Yves Jaffray pour les discussions qui m'ont aidé dans la préparation de cet article et les relecteurs anonymes pour leurs remarques sur une version antérieure qui ont contribué à le rendre plus clair.

Annexe

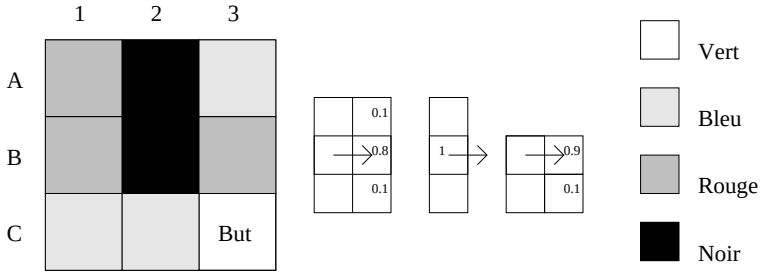

Figure 1. *Exemple de problème avec des préférences qualitatives*

Nous allons montrer le fonctionnement de l'algorithme 4.1.2 sur une instance très simple (fig. 1) du problème de navigation du robot dans un environnement hostile. Les états du PDMG sont les positions dans la grille. Les actions possibles du robot sont (N)ord, (S)ud, (E)st, (O)uest. Au centre de la figure 1, nous donnons les transitions probabilistes du robot quand il se déplace vers l'est. Les transitions des autres actions se déduisent par symétrie et rotation. L'ensemble des coûts qualitatifs sont ceux décrits dans le paragraphe précédent. La fonction de coûts est indiquée dans la figure 1. Par exemple, au déplacement de la case A1 dans la case A2 est affecté le coût Noir.

Nous reprenons les relations de préférence décrites ci-dessus (section 4.1.3) : relation de Bossong *et al.* (1997) pour les historiques et relation de dominance stochastique généralisée pour les loteries. Ainsi, la relation de préférence sur les politiques est entièrement déterminée.

Nous souhaitons que le robot atteigne la case but. Cette case atteinte, on suppose que le robot stationne. De plus, pour éviter les mouvements vers les murs, qui ne déplacent pas le robot, on suppose que pour ces actions, il existe un niveau de risque qui permet de les considérer pire que toute autre action.

A l'étape 1, dans la case A1, les actions Nord et Ouest ne déplacent pas le robot. On ne calcule donc pas leurs loteries associées. A l'action Est correspond la loterie $L^E_{A1,1} = (1/\mathrm{N})$. Les indices dans cette notation sont simplement : l'action (E), l'état (A1) et l'étape (1). Pour l'action Sud, on définit la loterie $L^S_{A1,1} = (0.1/\mathrm{N}, 0.9/\mathrm{R})$. L'action Sud est donc optimale.

De la même façon, dans la case A2, on obtient les loteries suivantes : $L^E_{A2,1} = (0.1/\text{R}, 0.9/\text{B})$ pour Est, $L^S_{A2,1} = (0.8/\text{N}, 0.2/\text{R})$ pour Sud, $L^O_{A2,1} = (1/\text{R})$ pour Ouest. On a $L^S_{A2,1} \prec_{DS} L^E_{A2,1}$, $L^S_{A2,1} \prec_{DS} L^O_{A2,1}$ et $L^O_{A2,1} \prec_{DS} L^E_{A2,1}$. Par conséquent, la meilleure action est Est.

Dans la cas A3, on obtient les loteries suivantes : $L^S_{A3,1} = (0.1/\text{N}, 0.9/\text{R})$ pour Sud, $L^O_{A3,1} = (1/\text{N})$ pour Ouest. Par conséquent, la meilleure action est Sud.

Dans la cas B1, on obtient les loteries suivantes : $L^E_{B1,1} = (0.9/\text{N}, 0.1/\text{B})$ pour Est, $L^S_{B1,1} = (1/\text{B})$ pour Sud, $L^N_{B1,1} = (0.1/\text{N}, 0.9/\text{R})$ pour Nord. Par conséquent, la meilleure action est Sud.

Pour la cas B2, l'action Sud est optimale avec la loterie $L^S_{B2,1} = (0.9/\text{B}, 0.1/\text{V})$. Pour la case B3, l'action Sud est optimale avec la loterie $L^S_{B3,1} = (0.1/\text{B}, 0.9/\text{V})$. Pour la case C1, l'action Est est optimale avec la loterie $L^E_{C1,1} = (0.1/\text{N}, 0.9/\text{B})$. Pour la case C2, la meilleure action est Est avec la loterie $L^E_{C2,1} = (0.1/\text{R}, 0.9/\text{V})$.

A l'étape 1, il y a une seule règle de décision optimale (fig. 2 pour notre structure de préférence.

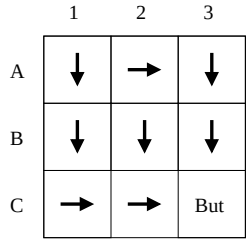

Figure 2. *Règle de décision optimale à l'étape 1*

A l'étape 2, dans la case A1, les actions possibles sont les suivantes : Est ou Sud. La loterie induite par l'action Est est égale à $L^E_{A1,2} = (0.9/L^E_{A2,1}, 0.1/L^S_{B2,1}) = (0.09/\text{NR}, 0.9/\text{NB}, 0.01/\text{NV})$. La loterie pour l'action Sud est égale à $L^S_{A1,2} = (0.9/L^S_{B1,1}, 0.1/L^E_{B2,1})$ qui se réécrit en $L^S_{A1,2} = (0.09/\text{NB}, 0.01/\text{NV}, 0.9/\text{RB})$. Par conséquent, l'action optimale est Sud.

Pour la case A2, l'action Est induit la loterie suivante $L^E_{A2,2}$ qui a pour valeur $(0.09/\text{NB}, 0.82/\text{RB}, 0.09/\text{RV})$. Pour l'action Sud, la loterie induite s'écrit $L^S_{A2,2} = (0.72/\text{NB}, 0.08/\text{NV}, 0.11/\text{RB}, 0.09/\text{RV})$. Enfin la loterie associée à Ouest vaut $L^O_{A2,2} = (0.09/\text{NR}, 0.81/\text{RR}, 0.1/\text{RB})$. L'action optimale est donc Est.

Pour la case A3, la meilleure action est Sud. Sa loterie associée est égale à $L^S_{A3,2} = (0.09/\text{NB}, 0.01/\text{NV}, 0.09/\text{RB}, 0.81/\text{RV})$.

Pour la case B1, la meilleure action est Sud. Sa loterie associée est égale à $L^S_{B1,2} = (0.09/\text{NB}, 0.01/\text{RB}, 0.81/\text{BB}, 0.09/\text{BV})$.

Pour la case B2, l'action optimale est Sud. Sa loterie associée est égale à $L^S_{B2,2} = (0.01/\text{NB}, 0.08/\text{RB}, 0.09/\text{BB}, 0.72/\text{BV}, 0.1/\text{V})$.

Pour la case B3, l'action optimale est Sud. Sa loterie associée est égale à $L^S_{B3,2} = (0.01/\text{RB}, 0.09/\text{BV}, 0.9/\text{V})$.

Pour la case C1, l'action optimale est Est. Sa loterie associée est égale à $L^E_{C1,2} = (0.09/\text{NB}, 0.01/\text{NV}, 0.09/\text{RB}, 0.81/\text{BV})$.

Pour la case C2, l'action optimale est Est. Sa loterie associée est égale à $L^E_{C2,2} = (0.01/\text{RB}, 0.09/\text{RV}, 0.9/\text{V})$.

A l'horizon 2, il y a donc une règle de décision optimale, qui est d'ailleurs identique à celle de l'horizon 1. On pourrait procéder de même pour les horizons supérieures.

6. Bibliographie

Bonet B., Pearl J., « Qualitative MDPs and POMDPs : An order-of-magnitude approximation », *UAI*, vol. 18, p. 61-68, 2002.

Bossong U., Schweigert D., Minimal paths on ordered graphs, Technical report, Report in Wirtschaftsmathematik no. 24/1997, University of Kaiserslautern, 1997.

Cavazos-Cadena R., de Oca R. M., « Nearly optimal policies in risk-sensitive positive dynamic programming on discrete spaces », *Mathematical Methods of Operations Research*, vol. 52, p. 133-167, 2000.

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.

Furukawa N., « Vector-valued markovian decision processes with countable state space », *Ann. Math. Stat.*, 1965.

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

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

Henig M., « Vector-valued dynamic programming », *SIAM Journal of control and optimization*, vol. 3, p. 490-499, 1983.

Krantz D., Luce R., Suppes P., Tversky A., *Foundations of measurement*, vol. Additive and Polynomial Representations, Academic Press, 1971.

Krebs D., Porteus E., « Dynamic choice theory and dynamic programming », *Econometrica*, vol. 47, n° 1, p. 91-100, 1979.

Machina M. J., « Dynamic consistency and non-expected utility models of choice under uncertainty », *Journal of Economic Literature*, vol. 27, n° 4, p. 1622-1668, 1989.

Novák J., « Linear programming in vector criterion markov and semi-markov decision processes », *Optimization*, vol. 20, p. 651-670, 1989.

P. Perny O. S., « An axiomatic approach to robustness in search problems with multiple scenarios », *UAI*, vol. 19, p. 469-476, 2003.

Perny P., Spanjaard O., « On preference-based Search in State Space Graphs », *AAAI*, vol. 14, p. 751-756, 2002.

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.

Spanjaard O., Exploitation de préférences non-classiques dans les problèmes combinatoires : modèles et algorithmes pour les graphes, PhD thesis, Université Paris IX Dauphine, 2003.

Viswanathan B., Aggarwal V., Nair K., « Multiple criteria markov decision processes », *TIMS Studies in the management sciences*, vol. 6, p. 263-272, 1977.

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

Wakuta K., « Optimal stationary policies in the vector-valued Markov decision process », *Stochastic processes and their applications*, vol. 42, p. 149-156, 1992.

Wakuta K., « Vector-valued markov decision processes and the systems of linear inequalities », *Stochastic processes and their applications*, vol. 56, p. 159-169, 1995.

White D., « Multi-objective infinite-horizon discounted markov decision processes », *Journal of mathematical analysis and applications*, vol. 89, p. 639-647, 1982.

Wilson N., « An order of magnitude calculus », *UAI*, vol. 11, p. 548-555, 1995.

Yu S. X., Lin Y., Yan P., « Optimization models for the first arrival target distribution function in discrete time », *Journal of mathematical analysis and applications*, vol. 225, p. 193-223, 1998.