Processus décisionnels de Markov : des récompenses ordinales au multicritère

Paul Weng

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

RÉSUMÉ. *Le modèle des processus décisionnels de Markov (MDP) offre un cadre général pour la résolution de problèmes de décision séquentielle dans l'incertain. Son exploitation suppose une connaissance précise des valeurs des paramètres (probabilités et récompenses). Dans ce papier, les récompenses sont qualitatives ou ne sont connues que de manière imparfaite. Seul un ordre est supposé connu. Un MDP à récompenses ordinales (OMDP) peut être vu comme un MDP à récompenses numériques vectorielles dans lequel les fonctions de valeur se transforment en distributions de probabilité. Nous listons alors quelques critères d'optimalité provenant d'ordres sur les distributions de probabilité, notamment la dominance probabiliste à points de référence. Les OMDP exploitant cette dominance sont équivalents à des MDP multicritères avec une priorité définie sur les critères. Pour ces derniers, à l'horizon infini un nouvel algorithme de résolution est proposé quand la priorité sur les critères est un préordre complet.*

ABSTRACT. *The model of Markov decision processes (MDP) is a general framework for solving sequential decision-making problems under uncertainty. Its exploitation assumes a perfect knowledge of the parameter values (probabilities and rewards). In this paper, rewards are either qualitative or only known imperfectly. Only an order is assumed to be known. An MDP with ordinal rewards (OMDP) can be viewed as an MDP with vectorial numeric rewards in which value functions can be transformed into probability distributions. We then list several optimality criteria coming from orders over probability distributions, notably the probabilistic dominance with reference points. OMDPs exploiting this dominance are equivalent to multicriteria MDPs with a priority defined on criteria. For such multicriteria MDPs, we propose for the infinite horizon a new solving algorithm when the priority over criteria is a total preorder.*

MOTS-CLÉS : *processus décisionnel de Markov, récompense qualitative, MDP multicritère.*

KEYWORDS: *Markov decision process, qualitative reward, multicriteria MDP.*

RSTI - RIA – 24/2010. Formalismes et modèles en PDA, pages 505 à 524

1. Introduction

Le modèle des processus décisionnels de Markov (MDP) est un modèle très étudié en intelligence artificielle (Sigaud *et al.,* 2008). Il offre un formalisme pour modéliser et résoudre des problèmes de planification dans l'incertain.

Son exploitation peut parfois poser problème dans la mesure où il nécessite la connaissance précise de ses paramètres (probabilités de transition et récompenses). Dans de nombreuses situations réelles, ceux-ci ne sont connus que de manière imprécise car il peut être difficile ou coûteux, voire impossible, de les déterminer de manière exacte. Les solutions d'un MDP pouvant bien entendu être très sensibles à ces paramètres, souvent il est délicat de fixer ses valeurs.

Ce constat a motivé des travaux récents sur les MDP robustes (Givan *et al.,* 2000, Bagnell *et al.,* 2001, Nilim *et al.,* 2003, Trevizan *et al.,* 2007). Dans ces travaux, les probabilités seules ou les probabilités avec les récompenses sont supposées connues de manière incertaine. Les paramètres peuvent alors être représentés par des intervalles au lieu de valeurs précises par exemple. Mais généralement, cette direction de recherche s'est focalisée principalement sur la méconnaissance des probabilités.

Les MDP possibilistes (Sabbadin, 1998) permettent de prendre en compte les situations où les récompenses sont qualitatives. Toutefois, la représentation de l'incertain est alors forcément possibiliste, ce qui peut être inadapté quand les probabilités de transition sont connues. À notre connaissance, il n'existe pas d'extension des MDP permettant à la fois une représentation probabiliste de l'incertain et une représentation qualitative ou ordinale des récompenses.

Dans ce papier, nous nous intéressons au cas où seules les récompenses sont qualitatives ou mal connues. Ainsi nous supposons qu'il est possible de déterminer les valeurs des probabilités. Nous pensons que ce cas est relativement courant. En effet, s'il existe généralement des moyens pour évaluer les probabilités de transition d'un système (expérimentation et estimation statistique par exemple), les préférences imposées par les valeurs de récompense sont souvent plus délicates à définir. Sur quelles bases théoriques choisirait-on de fixer une récompense à telle valeur plutôt qu'à telle autre ? Les travaux en théorie du mesurage (Krantz *et al.,* 1971) suggèrent que cette tâche est difficile. C'est pourquoi nous supposons dans notre cadre de travail que seul un ordre est connu sur les différentes récompenses possibles.

Nous montrons dans ce papier qu'un tel MDP avec des récompenses ordinales peut être interprété comme un MDP particulier avec des récompenses numériques vectorielles (VMDP). Dans ce VMDP, les fonctions de valeur peuvent alors être identifiées à des distributions de probabilité sur l'ensemble des récompenses ordinales. Il est alors naturel d'importer dans notre cadre des relations d'ordre connues sur les distributions de probabilité – nous présenterons entre autres la dominance stochastique du premier ordre (Shaked *et al.,* 1994) et la dominance probabiliste à points de référence (Castagnoli *et al.,* 1996) – pour définir la notion de «bonne» politique. Nous nous intéressons plus particulièrement à la dominance probabiliste avec plusieurs points de référence

(où une priorité est définie sur les points de référence). Avec cette relation, un MDP ordinal est équivalent à un nouvel MDP multicritère où à un critère correspond un point de référence. À l'horizon fini, nous rappelons qu'un algorithme d'induction arrière permet de déterminer les politiques préférées. À l'horizon infini, nous proposons un nouvel algorithme de résolution, fondé sur la programmation linéaire multiobjectif dans le cas où la priorité sur les critères est un préordre [1] complet.

Dans la section suivante, nous rappelons brièvement le modèle des processus décisionnels de Markov. Nous montrons qu'à partir de trois récompenses différentes possibles, la détermination de leurs valeurs peut devenir problématique. Dans cette section, nous introduisons également les MDP à récompenses ordinales et les VMDP associées. Nous montrons dans quelle mesure les fonctions de valeur d'un tel VMDP peuvent être identifiées à des distributions de probabilité sur les récompenses. Dans la section 3, nous passons en revue quelques ordres qui pourraient être envisageables sur ces distributions de probabilité. Nous présentons notamment la dominance probabiliste à points de référence. Dans la section 4, nous proposons les méthodes de résolution associées.

2. Cadre de travail

2.1. *Processus décisionnels de Markov*

Le modèle des *processus décisionnels de Markov* (MDP) se définit généralement par la donnée d'un quadruplet (S, A, p, r) (Sigaud *et al.,* 2008) :

– S un ensemble d'états,

– A un ensemble d'actions,

– $p\colon S \times A \to \mathcal{P}(S)$ une fonction de transition où $\mathcal{P}(S)$ est l'ensemble des distributions de probabilité sur S,

– $r\colon S \times A \to X \subset \mathbb{R}$ une fonction de récompense.

Dans notre cadre de travail, l'ensemble des états S et l'ensemble des actions A sont supposés finis. La fonction de transition fournit les probabilités d'occurrence des états futurs après l'exécution d'une action dans un état. Suivant la coutume, nous noterons $p(s, a, s') = p(s, a)(s')$ la probabilité d'atteindre l'état s' après l'exécution de l'action a dans l'état s. La fonction de récompense donne la récompense immédiate que reçoit l'agent après avoir exécuté une action dans un état. L'ensemble X – ensemble fini car S et A sont finis – représente l'ensemble de toutes les récompenses possibles.

Une *règle de décision* δ indique quelle action choisir dans chaque état à une étape donnée. Elle peut être *pure* : $\delta : S \to A$ est alors une fonction de l'ensemble des états S dans l'ensemble des actions A. Elle peut être également dite *mixte* : $\delta : S \to \mathcal{P}(A)$ est alors une fonction de l'ensemble des états S dans l'ensemble des distributions de

1. Un préordre est une relation binaire symétrique et transitive.

probabilité sur les actions A. Le choix de l'action à effectuer dans un état est donc choisi aléatoirement selon la distribution de probabilité. Remarquons qu'une règle de décision pure est une règle mixte dont les distributions de probabilité sont dégénérées (une seule action ayant une probabilité de 1 pour chaque état). Une *politique* π à un horizon h est une séquence de h règles de décision, notée $\pi = (\delta_1, \ldots, \delta_h)$ où chaque δ_i est une règle de décision. Elle est dite *pure* quand elle ne contient que des règles de décision pures et *mixte* autrement. L'ensemble des politiques pures (resp. mixtes) à l'horizon h sera noté Π_h^P (resp. Π_h^M). À l'horizon infini, une politique est simplement une séquence infinie de règles de décision. L'ensemble de ces politiques pures (resp. mixtes) sera noté Π_∞^P (resp. Π_∞^M). Une politique est dite *stationnaire* si à chaque étape de décision, la même règle de décision est utilisée.

Une *relation de préférence* $\succsim$, qui est simplement une relation binaire, est supposée définie sur les politiques. Une politique π *domine* ou est dite *préférée à* une autre politique π' quand $\pi \succsim \pi'$. On écrira $\pi \succ \pi'$ (resp. $\pi \sim \pi'$) quand $\pi \succsim \pi'$ et non $\pi' \succsim \pi$ (resp. $\pi \succsim \pi'$ et $\pi' \succsim \pi$). Une politique π est dite *non dominée* ou *préférée* quand il n'existe aucune politique π' telle que $\pi' \succ \pi$. Elle est dite *optimale* quand elle est préférée à toute autre politique. Remarquons qu'une politique optimale n'existe pas forcément car la relation de préférence pouvant n'être que partielle dans le cas général, certaines politiques peuvent être incomparables. La résolution d'un MDP consiste à déterminer une politique préférée pour un certain système de préférence. Nous présentons maintenant comment ces préférences sont définies dans le cadre classique.

La trace de l'exécution d'une politique est appelée *historique*. Les historiques, qui peuvent être de longueur finie ou infinie, débutant de l'état $s_0 \in S$ correspondent aux séquences suivantes : $(s_0, a_1, s_1, a_2, s_2, \ldots)$ où $\forall i \in \mathbb{N}, (a_i, s_i) \in A \times S$. La valeur d'un historique $\gamma = (s_0, a_1, s_1, a_2, s_2, \ldots, a_h, s_h)$ peut être définie de plusieurs manières. On peut simplement sommer les récompenses obtenues tout au long de l'historique :

$$r_t(\gamma) = \sum_{i=0}^{h-1} r(s_i, a_{i+1})$$

On peut également faire la somme actualisée des récompenses :

$$r_\beta(\gamma) = \sum_{i=0}^{h-1} \beta^i r(s_i, a_{i+1})$$

où $\beta \in [0, 1[$ est le facteur d'actualisation. Enfin, il est possible de considérer la moyenne des récompenses :

$$r_m(\gamma) = \frac{1}{h} \sum_{i=0}^{h-1} r(s_i, a_{i+1})$$

Ces valeurs peuvent s'étendre quand l'historique est de longueur infinie. Pour la somme des récompenses (sans facteur d'actualisation), il peut y avoir des problèmes de convergence. La somme actualisée est bien définie grâce au facteur d'actualisation.

Pour la moyenne des récompenses, la valeur de l'historique est définie comme une limite quand elle existe :

$$r_m(\gamma) = \lim_{h\to\infty} \frac{1}{h} \sum_{i=0}^{h-1} r(s_i, a_{i+1})$$

Une règle de décision δ depuis un état initial s induit une distribution de probabilité sur les historiques (de longueur 1). Comme on peut associer une valeur à tout historique, elle induit également une *loterie* (i.e. une distribution de probabilité) sur l'ensemble X des récompenses possibles. Cette loterie est égale à $p(s, \delta(s))$. Par induction, à une politique π dans un état initial s donné, on peut associer une distribution de probabilité sur les historiques. Ainsi, une politique induit également une loterie sur les valeurs des historiques. Par conséquent, il est possible de définir l'espérance des récompenses que peut générer une politique π dans un état s.

La fonction $v^\pi : S \to \mathbb{R}$ qui associe à chaque état s l'espérance des récompenses que peut obtenir une politique π est appelée la *fonction de valeur* de π :

$$v^\pi(s) = E_s^\pi(r_*(\Gamma))$$

où $* \in \{t, \beta, m\}$ et Γ est une variable aléatoire sur les historiques. Une politique π sera dite *préférée* à une autre politique π' si et seulement si :

$$\pi \succsim \pi' \Leftrightarrow \forall s \in S, v^\pi(s) \geq v^{\pi'}(s)$$

Dans ce cadre classique, la relation de préférence ainsi définie garantit l'existence d'une politique stationnaire pure optimale.

Selon qu'on utilise la somme, la somme actualisée ou la moyenne des récompenses, on dit que la fonction de valeur repose sur le *critère total*, le *critère total pondéré* ou le *critère moyen*. Ces trois critères sont fondés sur l'espérance, la différence résidant dans la manière dont sont valuées les historiques. En fait, ces critères sont des instances d'*utilités espérées* (Bouyssou *et al.,* 2000) où les valeurs des historiques jouent le rôle des utilités. Ainsi, ce sont des utilités espérées pour lesquelles on suppose que ces utilités sont additivement décomposables.

Pour rester simple, par la suite, nous ne considérerons que le critère total pondéré pour éviter tout problème à l'horizon infini. Notre approche pourrait bien entendu être étendue aux deux autres critères.

2.2. *Motivations de notre travail*

L'approche classique n'est pas discutable quand tous les paramètres du modèle sont numériques et connus avec précision. Toutefois quand on ne peut qu'ordonner les récompenses, il peut être problématique de fixer arbitrairement leurs valeurs car les politiques préférées peuvent avoir une grande sensibilité à ces valeurs. Le problème

de déterminer les valeurs des récompenses, appelé *élicitation des préférences* est un problème épineux étudié en théorie de la décision (Bouyssou *et al.,* 2000).

Le problème que nous considérons dans ce papier n'apparaît que quand les préférences sont un peu complexes, c'est-à-dire quand on a besoin d'au moins trois valeurs distinctes pour définir la fonction de récompense. Quand deux valeurs suffisent, comme c'est le cas par exemple dans les problèmes où il n'y a qu'un seul ensemble d'états but (tous identiques) et où tous les autres états sont considérés comme équivalents, la définition de la fonction de récompenses peut être quelconque car elle n'influe pas sur la ou les politiques optimales.

Proposition 2.1. *Soit $r_1, r_2, r'_1, r'_2 \in \mathbb{R}$ tels que $r_1 < r_2$ et $r'_1 < r'_2$. Soit un MDP (S, A, p, r) dont la fonction de récompense ne peut prendre que les deux valeurs possibles r_1 et r_2. Posons r' la fonction définie à partir de r en substituant r_1 par r'_1 et r_2 par r'_2. Alors les MDP (S, A, p, r) et (S, A, p, r') ont la ou les même politiques optimales.*

Démonstration. Les deux MDP possèdent les mêmes politiques. Il suffit de montrer que le sens des préférences entre deux politiques n'est pas modifié selon que l'on utilise r ou r'. On constate qu'il est possible de passer de l'une à l'autre par une transformée affine positive. Comme le critère total pondéré est linéaire, une telle transformation conserve les inégalités et donc les sens de préférence entre deux politiques. □

Pour le cas où la fonction de récompense nécessite trois valeurs distinctes, le choix de ces trois valeurs peut avoir un impact important sur les politiques optimales comme nous l'illustrons sur un exemple simple.

Exemple 2.1. *Considérons le MDP suivant où $S = \{1, 2\}$ et $A = \{a, b\}$. Le facteur d'actualisation est fixé à $\beta = 0.5$. La fonction de transition est définie comme suit :*

$$p(1, a, 1) = 1 \quad p(1, b, 1) = 0,5 \quad p(2, a, 1) = 1$$

Pour simplifier, on suppose que l'action b n'est pas possible dans l'état 2. Dans ce cas, il n'existe que deux politiques stationnaires déterministes selon le choix de l'action dans le premier état. Supposons que l'on sache seulement $r(1, b) > r(1, a) > r(2, a)$. Si la fonction de récompense est définie arbitrairement ainsi :

$$r(1, a) = 1 \quad r(1, b) = 2 \quad r(2, a) = 0$$

alors on vérifie aisément que la meilleure politique est celle qui consiste à choisir l'action b dans l'état 1. La fonction de valeur obtenue dans cet état vaut $\frac{16}{5} = 3,2$ contre 2 pour l'autre politique.

Maintenant, si la fonction de récompense avait été définie ainsi :

$$r(1, a) = 9 \quad r(1, b) = 10 \quad r(2, a) = 0$$

La meilleure politique aurait été celle qui choisit l'action a. Sa fonction de valeur en l'état 1 vaudrait 18 contre 16 pour l'autre politique.

Bien que les deux fonctions respectent l'ordre imposé sur les récompenses, on observe une inversion des préférences. Ainsi le choix de l'échelle de valuation des récompenses peut être déterminant sur la politique optimale, ce qui peut être problématique dans certaines situations.

2.3. *MDP à récompenses ordinales*

Comme les préférences sur les politiques peuvent être sensibles au choix des valeurs des récompenses, nous proposons dans ce papier de ne pas introduire arbitrairement cette information quand elle n'est pas connue. Dans les situations où l'on ne possède qu'une information ordinale sur les récompenses, un modèle semi-qualitatif des MDP peut être exploité. La fonction de récompenses $r : S \times A \rightarrow E$ est alors définie sur une échelle qualitative $(E, >)$ complètement ordonnée, le nombre de pas de cette échelle étant le nombre de valeurs différentes de récompenses dont on a besoin pour modéliser les préférences du problème considéré. L'échelle est nécessairement finie car les ensembles S et A sont supposés finis.

Ce MDP semi-qualitatif peut être transformé en un MDP particulier avec des récompenses vectorielles (VMDP). Soit $n \in \mathbb{N}$ le nombre de pas de l'échelle $E = \{r_1 > r_2 \ldots > r_n\}$. Définissons le vecteur à n dimensions $1_i = (0, \ldots, 0, 1, 0, \ldots, 0)$, nul partout sauf à la i-ème position où il vaut 1. Pour un vecteur x, nous notons $x^{(i)}$ sa i-ème composante. Le MDP semi-qualitatif peut être vu comme un VMDP où la fonction de récompense $\hat{r} : S \times A \rightarrow \mathbb{R}^n$ est définie à partir de r par $\forall s \in S, \forall a \in A, \hat{r}(s, a) = 1_i$ si $r(s, a) = r_i$. Ce VMDP est un MDP multicritère classique (Viswanathan *et al.,* 1977) dans lequel on fait l'hypothèse supplémentaire qu'il existe une préférence sur les critères.

Comme pour un MDP multicritère classique, dans ce VMDP, on peut définir récursivement la fonction de valeur $\hat{v}_h^\pi$ d'une politique π à un horizon h par :

$$\hat{v}_0^\pi(s) = (0, \ldots, 0) \in \mathbb{R}^n \quad \forall s \in S \qquad [1]$$

$$\hat{v}_t^\pi(s) = \hat{r}(s, \delta_t(s)) + \beta \sum_{s' \in S} p(s, \delta_t(s), s') \hat{v}_{t-1}^\pi(s') \quad \forall s \in S, \forall t > 0 \qquad [2]$$

où $\pi = (\delta_1, \ldots, \delta_h)$ et les sommes et produits sur les vecteurs sont calculés composante par composante. De plus, ces fonctions de valeur sont également bien définies à l'horizon infini grâce au facteur $\beta \in [0, 1[$. Elles seront simplement notées $\hat{v}^\pi$.

Avec ces réécritures, le problème de définir une relation de préférence dans le MDP semi-qualitatif revient à définir une relation de préférence sur des vecteurs, permettant ainsi de comparer les fonctions de valeur et donc leurs politiques associées. Remarquons que bien qu'on ait affaire à des vecteurs de valuation, la dominance de Pareto [2] n'a pas de sens ici puisqu'il existe une relation de préférence sur les critères.

2. Un vecteur x Pareto-domine un vecteur y si et seulement si pour tout i, $x^{(i)} \geq y^{(i)}$ et il existe j, $x^{(j)} > y^{(j)}$

En effet, on préférera par exemple le vecteur 1_i à 1_j quand $i < j$ car il représente une récompense meilleure. Dans la section suivante, nous proposons et passons en revue quelques relations potentiellement intéressantes. Au passage, notons que si finalement les récompenses sont numériques et connues, il est possible de faire le lien entre les fonctions de valeur du MDP multicritère et du MDP classique.

Proposition 2.2. *On a :*

$$\forall t > 0, \forall s \in S, v_t^\pi(s) = \sum_{i=1}^{n} \hat{v}_t^\pi(s)^{(i)} r^{(i)}$$

et

$$\forall s \in S, v^\pi(s) = \sum_{i=1}^{n} \hat{v}^\pi(s)^{(i)} r^{(i)}$$

3. Relations de préférence

3.1. *Ordre lexicographique*

Du fait de l'ordre sur l'échelle E, une première idée, naturelle et simple, serait de comparer les vecteurs de valeur selon l'ordre lexicographique suivant :

$$x \succsim y \Leftrightarrow \exists i = 1, \ldots n, \begin{cases} \forall j < i, x^{(j)} = y^{(j)} \\ \text{et } x^{(i)} > y^{(i)} \end{cases} \quad [3]$$

Son interprétation dans notre cadre est simple. Dans la comparaison de deux fonctions de valeur dans un état, on souhaite que la première composante soit la plus élevée possible car elle correspond à la meilleure récompense. S'il y a égalité, on s'intéressera à la seconde composante et ainsi de suite.

Toutefois, l'inconvénient avec l'ordre lexicographique est qu'on interdit les compensations entre récompenses. Ainsi $(1, 0, \ldots, 0, 0)$ sera préféré à $(0, 100, \ldots, 100, 100)$, ce qui peut être discutable. Par ailleurs, on peut vouloir avoir un degré plus élevé d'expressivité dans la définition des préférences.

3.2. *Relations sur les distributions de probabilité*

Nous allons montrer dans cette section que les fonctions de valeur en un état dans le VMDP peuvent être identifiées à des distributions de probabilité. En faisant ce lien, il sera alors naturel de vouloir comparer les fonctions de valeur en utilisant des relations de préférence sur les distributions de probabilité.

En un état donné, le vecteur associé à une fonction de valeur est substantiellement une distribution de probabilité. En effet, à l'horizon fini h, si on divise la fonction de valeur $\hat{v}_h^\pi$ de la politique π par la valeur u_h définie par :

$$u_h = \sum_{t=0}^{h} \beta^t \quad [4]$$

qui représente la somme des poids utilisés dans la pondération des récompenses, on obtient le résultat suivant :

Proposition 3.1. $\forall h > 0, \forall \pi \in \Pi_h^M, \forall s \in S, \frac{\hat{v}_h^\pi(s)}{u_h}$ *est une distribution de probabilité sur* E.

Démonstration. La démonstration se fait simplement par récurrence sur h. Pour $h = 1$, $\hat{v}_1^\pi(s) = \hat{r}(s, \pi_1(s))$ est une distribution de probabilité (dégénérée) sur E. Maintenant, supposons que $\frac{\hat{v}_h^\pi(s)}{u_h}$ soit une distribution de probabilité pour tout $s \in S$. D'après l'équation 2, on a :

$$\begin{aligned}
\frac{\hat{v}_{h+1}^\pi(s)}{u_h + 1} &= \frac{\hat{r}(s, \pi_{h+1}(s)) + \beta \sum_{s' \in S} p(s, \pi_{h+1}(s), s') \hat{v}_h^\pi(s)}{u_{h+1}} \\
&= \frac{\hat{r}(s, \pi_{h+1}(s)) + \beta \sum_{s' \in S} p(s, \pi_{h+1}(s), s') u_h \frac{\hat{v}_h^\pi(s)}{u_h}}{u_{h+1}} \\
&= \frac{\hat{r}(s, \pi_{h+1}(s)) + \beta u_h \sum_{s' \in S} p(s, \pi_{h+1}(s), s') \frac{\hat{v}_h^\pi(s)}{u_h}}{u_{h+1}}
\end{aligned}$$

On constate donc que $\hat{v}_{h+1}^\pi(s)/(u_h + 1)$ s'écrit comme combinaison linéaire (avec des poids positifs sommant à 1) de distributions de probabilité et est donc également une distribution de probabilité. □

Cette distribution de probabilité peut s'interpréter comme la proportion de chacune des récompenses dans la fonction de valeur en un état donné. À l'horizon infini, on observe simplement que $\frac{\hat{v}_t^\pi(s)}{u_t}$ converge vers une distribution de probabilité.

Proposition 3.2. *Pour toute politique* $\pi \in \Pi_\infty^M$*, pour tout état* $s \in S$*, on a :*

$$\lim_{t \to \infty} \frac{\hat{v}_t^\pi(s)}{u_t} = (1 - \beta)\hat{v}^\pi(s)$$*, qui est une distribution de probabilité sur* E.

Démonstration. Soit une politique $\pi \in \Pi_\infty^M$. Comme on suppose que $\beta \in [0, 1[$, pour tout $s \in S$, la suite $\hat{v}_t^\pi(s)$ converge vers $\hat{v}^\pi(s)$. La suite u_t est une série géométrique de raison β. Elle converge donc vers $\frac{1}{1-\beta}$. Par conséquent, $\frac{\hat{v}_t^\pi(s)}{u_t}$ converge vers $(1 - \beta)\hat{v}^\pi(s)$. Comme la somme des composantes du vecteur $\frac{\hat{v}_t^\pi(s)}{u_t}$ vaut 1 pour tout t, cette propriété est conservée après passage à la limite. Par ailleurs, toutes les valeurs de $\hat{v}^\pi(s)$ étant bien entendu positives, $(1-\beta)\hat{v}^\pi(s)$ est bien une distribution de probabilité. □

Étant donné que comparer deux fonctions de valeur revient à comparer les distributions de probabilité associées (u_t et $1-\beta$ étant des constantes), il est alors naturel de vouloir utiliser des relations de préférence sur les distributions de probabilité et de les importer dans notre cadre. Pour simplifier, nous noterons ces relations de préférence $\succsim$ également. Le contexte dira s'il s'agit d'une relation de préférence sur les distributions de probabilité ou celle sur les politiques. Nous commençons la présentation par la relation de préférence qui est la plus naturelle et la plus évidente : la dominance stochastique du premier ordre.

3.3. *Dominance stochastique du premier ordre*

La *dominance stochastique du premier ordre* se définit comme suit dans notre cadre, pour toute paire de distributions de probabilité P, P' sur E :

$$P \succsim P' \Leftrightarrow \forall x \in E, \sum_{y \geq x} P(y) \geq \sum_{y \geq x} P'(y)$$

Cette relation a une interprétation naturelle. Elle dit que pour toute récompense, la probabilité d'obtenir au moins cette récompense est plus grande pour la distribution de probabilité préférée. Malheureusement, cette dominance stochastique est généralement peu discriminante car elle est une relation de préférence partielle du fait de la condition "pour tout x". On peut vouloir alors la raffiner par une relation de préférence plus discriminante et/ou complète. Dans les sections suivantes, nous en présentons quelques-unes.

Au passage, notons que les critères d'utilité espérée (et notamment l'espérance) induisent des relations de préférence raffinant cette dominance. Ainsi les critères généralement utilisés dans les MDP (critère total, pondéré, moyen) sont également compatibles avec cette dominance du fait qu'ils reposent sur l'espérance.

3.4. *Dominance probabiliste*

Une idée naturelle pour raffiner la dominance stochastique du premier ordre est alors de vouloir comparer deux distributions de probabilité par la relation suivante, pour deux distributions de probabilité P, P' sur E :

$$P \succsim P' \Leftrightarrow \mathrm{P}(P \geq P') \geq \mathrm{P}(P' \geq P)$$

où $\mathrm{P}(P \geq P')$ est la probabilité que P obtienne de meilleures récompenses que P' et est définie par :

$$\mathrm{P}(P \geq P') = \sum_{x \in E} P(x) \sum_{y \leq x} P'(y).$$

Cette relation de préférence s'interprète naturellement : la distribution P est préférée à P' si et seulement si la probabilité que P obtienne de meilleures récompenses que P' est supérieure à la probabilité de l'évènement inverse.

Malheureusement, bien que cette relation de préférence soit complète, elle n'est pas transitive. En effet, il est possible d'observer des cycles :

Exemple 3.1. *Nous reprenons un exemple de (Perny* et al., *1999). Supposons que* $E = \{r_1 > r_2 > r_3 > r_4 > r_5\}$. *Considérons les trois distributions de probabilités* P, P', P'' *sur* E *définies par :*

	r_1	r_2	r_3	r_4	r_5
P	0	$0,51$	0	0	$0,49$
P'	0	0	1	0	0
P''	$0,49$	0	0	$0,51$	0

On vérifie alors aisément que $P \succ P' \succ P'' \succ P$.

Par conséquent, cela exclut la possibilité d'utiliser cette relation de préférence. Toutefois, en restant proche de cette idée, il existe un moyen de s'en sortir en introduisant un point de référence.

3.5. *Dominance probabiliste avec point de référence*

La *dominance probabiliste avec point de référence* est proche de la relation précédente. La différence réside dans l'introduction d'un point de référence, c'est-à-dire d'une valeur fixée à l'avance, qui peut être une constante ou plus généralement une distribution de probabilité pour la comparaison des loteries. Dans un cadre général où l'incertain n'est pas supposé probabiliste, ce critère de décision a été axiomatisé par (Perny *et al.,* 2006).

Considérons un point de référence ψ qui est une distribution de probabilité sur E. Nous supposerons que ψ est indépendante des loteries qui nous intéressent. Rappelons que ψ peut être une constante s'il est une distribution de probabilité dégénérée. La dominance probabiliste $\succsim^{\psi}$ avec le point de référence ψ se définit ainsi :

$$P \succsim^{\psi} P' \Leftrightarrow \mathrm{P}(P \geq \psi) \geq \mathrm{P}(P' \geq \psi) \qquad [5]$$

où $\mathrm{P}(P \geq \psi)$ est la probabilité que P obtienne de meilleures récompenses que le point de référence et vaut :

$$\mathrm{P}(P \geq \psi) = \sum_{x \in E} P(x) \sum_{y \leq x} \psi(y).$$

Intuitivement, cette dominance s'interprète de la manière suivante : quand on compare deux loteries, la préférée est celle dont la probabilité d'obtenir des récompenses meilleures que le point de référence est la plus grande. Quand le point de référence

est une constante, cette dominance revient simplement à comparer les probabilités d'obtenir au moins cette valeur de référence.

En fait, l'utilité espérée (et donc les critères total, total pondéré et moyen qui en sont des cas particuliers) peut être considérée comme un cas particulier de cette dominance, comme le soulignent (Castagnoli *et al.,* 1996) quand le point de référence est lui-même une distribution de probabilité. En effet, en posant $u(x) = \sum_{y \leq x} \psi(y)$, qui est simplement la fonction de répartition associée à ψ, on constate que l'équation 5 revient simplement à comparer des utilités espérées. Ce lien nous permet d'affirmer que cette dominance est transitive et définit bien un préordre et que de plus, elle raffine la dominance stochastique du premier ordre.

Dans notre cadre, le choix d'un point de référence ψ équivaut à la définition d'une fonction de récompense r^{ψ} définie par :

$$\forall s \in S, \forall a \in A, r^{\psi}(s,a) = u(r(s,a)) \quad [6]$$

Nous pouvons alors déduire que chercher les politiques optimales pour le MDP (S, A, p, r) avec cette dominance à point de référence est équivalent à résoudre le MDP classique (S, A, p, r^{ψ}) d'après la proposition 2.2.

Si finalement par le choix d'un point de référence, on fixe implicitement des valeurs pour les récompenses, de manière légitime, on peut s'interroger sur l'intérêt de notre cadre ordinal. Effectivement, le modèle de l'utilité espérée et le modèle de la dominance probabiliste à point de référence sont formellement équivalentes [3]. Cependant, la sémantique des deux modèles est très différente et cela a son importance dans notre cadre. En effet, dans les situations de connaissance imparfaite, le choix d'un point de référence permet une certaine justification et une interprétation naturelle de ce modèle de préférence que ne permet pas un choix arbitraire direct de valeurs numériques pour les récompenses.

Dans ce modèle de préférence, l'interprétation est simple : on cherche à maximiser la probabilité d'obtenir des récompenses meilleures qu'un point de référence. Par exemple, cette vision des choses permet de donner une interprétation simple pour les MDP classiques ayant des valeurs de récompenses espacées de manière régulière, c'est-à-dire si $r_1 > \ldots > r_n$ sont les différentes valeurs de récompenses possibles, alors il existe une constante C telle que $\forall i, r_i = r_{i+1} + C$. Pour cette classe de MDP, le modèle de la dominance probabiliste à point de référence nous indique que quand on fixe ainsi les valeurs des récompenses, on choisit implicitement comme point de référence la distribution de probabilité uniforme sur les différentes récompenses possibles [4]. L'interprétation est donc que dans ces MDP, les politiques optimales sont

3. En fait, le modèle de la dominance probabiliste à point de référence est plus générale que celui de l'utilité espérée car il pourrait permettre par ailleurs de relâcher l'hypothèse d'indépendance entre les loteries et le point de référence

4. Rappelons qu'une transformation affine positive des récompenses ne modifie pas les préférences sur les politiques. Ainsi, il est toujours possible de transformer les récompenses en une fonction de répartition.

celles qui maximisent la probabilité de faire mieux qu'un tirage aléatoire uniforme sur les récompenses. Ce modèle de préférence apporte un éclairage nouveau et intéressant aux MDP classiques ayant une fonction de récompense de cette forme.

Par ailleurs, l'interprétation de ce modèle de préférence est d'autant plus simple quand on choisit comme point de référence un pas de l'échelle E. Toutefois, dans ce cas, cette dominance bien que complète peut ne pas être très discriminante car on a une vision binaire des récompenses. Plus précisément, on utilise implicitement une fonction de récompense qui ne prend que deux valeurs possibles : 1 quand on est au dessus du point de référence et 0 sinon. On peut alors vouloir étendre cette dominance en introduisant plusieurs points de référence.

3.6. *Dominance probabiliste avec plusieurs points de référence*

Nous définissons cette *dominance à plusieurs points de référence* en nous inspirant d'une proposition de (Grosof, 1991, Junker, 2002) pour la généralisation de la dominance de Pareto dans le cadre multicritère.

Prenons un ensemble de k points de référence $\Psi = \{\psi_1, \ldots, \psi_k\}$. Soit $\rhd$ un ordre strict sur $\{1, \ldots, k\}$ qui définit une priorité sur les points de référence. Quand $i \rhd j$, on dit que le point de référence ψ_i est prioritaire sur ψ_j, ce qui sera également noté $\psi_i \rhd \psi_j$.

Cette famille de relations de préférence est définie de la manière suivante :

$$P \succ^{\Psi} P' \Leftrightarrow \left\{ \begin{array}{l} \exists i = 1, \ldots, k, P \succ^{\psi_i} P' \\ \forall i = 1, \ldots, k, P' \succ^{\psi_i} P \Rightarrow \exists j \rhd i, P \succ^{\psi_j} P' \end{array} \right. \quad [7]$$

La distribution P domine la distribution P' au sens de $\succ^{\Psi}$ si chaque fois que P' domine P pour un certain point de référence, P domine P' pour un point de référence plus prioritaire. La partie large de la relation se définit alors ainsi :

$$P \succsim^{\Psi} P' \Leftrightarrow \left\{ \begin{array}{l} P = P' \text{ ou} \\ P \succ^{\Psi} P' \end{array} \right. \quad [8]$$

Cette définition est très générale et englobe différentes relations de préférence sur les distributions de probabilité. En effet, la dominance stochastique du premier ordre en est un cas particulier quand tous les pas de l'échelle sont pris comme points de référence et que la relation $\rhd$ est vide. L'ordre lexicographique présenté dans la section 3.1 est une autre instance de cette dominance quand tous les pas de l'échelle sont pris en compte et que $\rhd$ est un ordre linéaire complet.

Cette dominance permet de formuler des préférences d'une grande expressivité. En effet, il est possible, par exemple, de définir une préférence du type : « maximiser la probabilité d'obtenir une récompense meilleure qu'une certaine récompense moyenne m et en cas d'égalité, maximiser la probabilité d'obtenir au moins une bonne récompense b » en prenant m au milieu de l'échelle E, b dans la partie supérieure de l'échelle et en donnant la priorité à m.

Dans la section précédente, nous avons vu qu'un point de référence définissait implicitement une certaine fonction de récompense. Ici, l'utilisation de plusieurs points de référence définit donc une fonction de récompense vectorielle. Ainsi, pour un ensemble de points de référence $\Psi = \{\psi_1, \ldots, \psi_k\}$, la fonction de récompense vectorielle implicitement définie, notée r^Ψ, est donnée par :

$$r^\Psi = (r^{\psi_1}, \ldots, r^{\psi_k})$$

De manière similaire, résoudre le MDP (S, A, p, r) (et donc le VMDP $(S, A, p, \hat{r})$) avec la dominance $\succsim^\Psi$ revient à résoudre le MDP multicritère (S, A, p, r^Ψ) avec une dominance avec priorité sur les vecteurs de récompense induite par $\succsim^\Psi$.

4. Méthode de résolution

Dans la section précédente, nous nous sommes intéressé à définir un système de préférence exploitable dans le cadre des MDP à récompenses ordinales. Comme nous l'avons vu, le cas où un seul point de référence est utilisé ne pose pas de problème car il est possible de le ramener à un MDP classique. Nous montrons maintenant comment calculer les politiques préférées au sens de la dominance probabiliste avec plusieurs points de référence.

À partir de maintenant, nous ne considérons que le MDP multicritère (S, A, p, r^Ψ) étant donné les équivalences précédentes. Nous notons $\overline{v} : S \to \mathbb{R}^k$ les fonctions de valeur induites par la fonction de récompense $r^\Psi : S \times A \to \mathbb{R}^k$. Par abus de notation, nous notons également $\succsim^\Psi$ la relation de préférence sur les vecteurs de $\mathbb{R}^k$. Comme la comparaison des distributions de probabilité se fait selon l'équation 7, les fonctions de valeur dans ce MDP multicritère en un état sont comparées selon la dominance à priorité sur les vecteurs définie par (Grosof, 1991, Junker, 2002) :

$$\overline{v}(s) \succ^\Psi \overline{v}'(s) \Leftrightarrow \begin{cases} \exists i, \overline{v}^{(i)}(s) > \overline{v}'^{(i)}(s) \\ \forall i, \overline{v}'^{(i)}(s) > \overline{v}^{(i)}(s) \Rightarrow \exists j \rhd i, \overline{v}^{(j)}(s) > \overline{v}'^{(j)}(s) \end{cases} \quad [9]$$

Nous noterons, pour un ensemble X, l'ensemble des éléments préférés ou non dominés de X : $M(X, \succsim) = \{x \in X : \forall y \in X, y \succsim x \Rightarrow y = x\}$.

4.1. *Horizon fini*

À l'horizon fini, un tel MDP multicritère a déjà été proposé par (Perny *et al.,* 2005). Les auteurs ont montré qu'un algorithme fondé sur l'induction arrière permet de déterminer les politiques préférées à l'horizon fini. Le système de préférence utilisé dans ce MDP multicritère est également un cas particulier du cadre étudié dans (Weng, 2006). Dans ces papiers, un algorithme d'induction arrière (figure 1) a été proposé.

Ici, du fait que la relation de priorité $\rhd$ sur les points de référence peut être partielle, l'algorithme travaille sur des ensemble de valeurs plutôt que sur des valeurs.

Figure 1. *Algorithme d'induction arrière généralisé*

$\forall s \in S, \overline{V}_0(s) \leftarrow \{(0, \ldots, 0)\}\,; t \leftarrow 0$
repeat
  $t \leftarrow t + 1$
  for all $s \in S$ **do**
    for all $a \in A$ **do**
      $\overline{Q}_t(s,a) \leftarrow r^{\Psi}(s,a) + \beta \sum_{s' \in S} p(s,a,s') \overline{V}_{t-1}(s')$
    end for
    $\overline{V}_t(s) \leftarrow M(\{\overline{Q}_t(s,a) : a \in A\}, \succsim^{\Psi})$
  end for
until $t = h$

Ainsi, les variables $\overline{V}_t(s)$ et $\overline{Q}_t(s,a)$ représentent des ensembles. En effet, l'opération de maximisation M fournit généralement un ensemble de valeurs plutôt qu'une meilleure valeur quand la relation $\rhd$ est partielle. Par ailleurs, il est nécessaire de conserver toutes les valeurs non dominées à chaque étape de calcul car une valeur non dominée à une étape donnée peut se révéler dominée finalement à une étape ultérieure.

4.2. *Horizon infini*

Le cas de l'horizon fini avait déjà été traité, ce qui n'est pas le cas de l'horizon infini. Nous supposons ici que la relation de priorité $\rhd$ induit un préordre complet et nous notons $\bowtie$ la relation d'équivalence associée ($\psi \bowtie \psi' \Leftrightarrow (\psi \rhd \psi'$ et $\psi' \rhd \psi)$). Cette hypothèse bien que simplificatrice permet encore de conserver de nombreuses relations. Par exemple, la dominance stochastique du premier ordre et l'ordre lexicographique respectent cette hypothèse.

4.2.1. *Programme linéaire multiobjectif*

La méthode de résolution que nous proposons repose sur le programme linéaire multiobjectif introduit par (Viswanathan *et al.,* 1977, Novák, 1989) pour la résolution de MDP multicritère (avec la dominance de Pareto). Nous le rappelons brièvement maintenant. Pour cela, donnons tout d'abord quelques définitions :

– La probabilité que le processus rentre dans l'état $s \in S$ et que l'action $a \in A$ soit exécuté à l'étape t est noté $p_s^a(t)$. Ces probabilités dépendent donc de la politique choisie. Si on définit une distribution de probabilité initiale $\mu = (\mu_1, \mu_2, \cdots, \mu_{|S|})$ sur les états, on a :

$$\begin{aligned} \sum_{a \in A} p_s^a(1) &= \mu_s, \forall s \in S \\ \sum_{a \in A} p_s^a(t) &= \sum_{s' \in S} \sum_{a \in A} p(s', a, s) p_{s'}^a(t-1), \forall s \in S, \forall t = 2, 3, \ldots \end{aligned} \qquad [10]$$

et par conséquent, $\sum_{s\in S}\sum_{a\in A} p_s^a(t) = 1, \forall t = 1, 2, \ldots.$

– Le critère total pondéré que l'on veut optimiser s'écrit :

$$\overline{v}^p(\mu) = \sum_{t=1}^{\infty} \beta^{t-1} \sum_{s\in S}\sum_{a\in A} p_s^a(t) r^{\Psi}(s,a)$$

où $p = (p_s^a(t))_{s\in S, a\in A, t>0}$.

– Définissons une variable intermédiaire nécessaire pour le programme linéaire.

$$x_s^a = \sum_{t=1}^{\infty} \beta^{t-1} p_s^a(t), \forall s \in S, \forall a \in A$$

La variable x_s^a peut être interprétée comme la fréquence actualisée d'être dans l'état s et de choisir l'action a.

Le programme linéaire multiobjectif s'écrit alors :

$$\begin{aligned} \text{v-max} \quad & \overline{v}(x) = \sum_{s\in S}\sum_{a\in A} x_s^a r^{\Psi}(s,a) \qquad [11] \\ \text{sous contraintes} \quad & \sum_{a\in A} x_s^a - \beta \sum_{s'\in S}\sum_{a\in A} p(s',a,s) x_{s'}^a = \mu_s, \forall s \in S \\ & x_s^a \geq 0, \forall s \in S, \forall a \in A \end{aligned}$$

où v-max est l'opérateur de maximisation vectorielle (au sens de la dominance de Pareto) et x est le vecteur composé des x_s^a. La première contrainte traduit simplement la relation 10. Nous noterons ce programme $PL(\Psi)$ pour indiquer qu'il est formulé avec la dominance de Pareto prenant en compte les points de référence de Ψ.

À chaque solution basique trouvée est associée une politique pure stationnaire non dominée (Viswanathan *et al.,* 1977, Novák, 1989) quand μ est une distribution de probabilité strictement positive. Retrouver la politique stationnaire non dominée à partir des x_s^a est alors très simple. En effet, pour un état s donné, un seul des x_s^a est non nul. Et il indique donc quelle action il faut choisir à l'état s.

De nombreuses méthodes ont été proposées pour résoudre les programmes linéaires multiobjectifs. On peut voir à ce sujet (Zeleny, 1974) ou (Steuer, 1986). Le plus simple est peut-être d'utiliser la généralisation de l'algorithme du simplexe. Nous ne détaillerons pas cette étape dans ce papier pour des raisons de concision.

4.2.2. *Méthode de résolution*

Avant de montrer comment déterminer les politiques préférées au sens de $\succsim^{\Psi}$, nous prouvons deux propositions qui nous sont utiles pour comprendre la méthode de résolution que nous énonçons ensuite. Mais donnons tout d'abord quelques définitions et notations. Appelons les K classes d'équivalence de $\bowtie$: $\Psi_1, \ldots, \Psi_K$ avec

$\forall \psi \in \Psi_i, \forall \psi' \in \Psi_j, \psi \rhd \psi'$ si $i < j$. Ainsi, pour une classe d'équivalence Ψ_i, $\succsim^{\Psi_i}$ est simplement la dominance de Pareto prenant seulement en compte les composantes correspondant aux points de référence de Ψ_i. Ici, les fonctions de valeur sont simplement des vecteurs de $\mathbb{R}^k$ en supposant qu'on ait fixé une distribution initiale μ. Pour deux vecteurs $\overline{v}, \overline{v}'$ et pour un ensemble quelconque de points de référence Ψ, nous rappelons que $\overline{v} \sim^{\Psi} \overline{v}'$ signifie que les valeurs de $\overline{v}$ et $\overline{v}'$ sont égales pour leurs composantes correspondant aux points de référence de Ψ.

Définissons maintenant les ensembles suivants :

$$M_0 = \{\overline{v}^\pi : \pi \in \Pi_\infty^M\} \quad [12]$$

$$M_1 = M(M_0, \succsim^{\Psi_1}) \quad [13]$$

$$M_i = \bigcup_{\overline{v} \in M_{i-1}} M(\{\overline{v}^\pi : \pi \in \Pi_\infty^M, \overline{v}^\pi \sim^{\cup_{k<i}\Psi_k} \overline{v}\}, \succsim^{\Psi_i}) \quad \forall i = 2 \ldots K \quad [14]$$

L'ensemble M_0 est l'ensemble des fonctions de valeur réalisables. L'ensemble M_1 est l'ensemble des fonctions de valeur non dominées pour les points de référence dans Ψ_1. L'ensemble M_i contient toutes les fonctions de valeur non dominées pour les points de référence dans Ψ_i et telles que leurs valeurs correspondant aux points de référence de Ψ_k sont égales à celles d'une fonction de valeur de M_k pour tout $k < i$. La première proposition fournit une propriété vérifiée par les politiques préférées au sens de $\succsim^{\Psi}$.

Proposition 4.1. *On a* $M(M_0, \succsim^{\Psi}) = M_K$.

Démonstration. Si $x \in M(M_0, \succsim^{\Psi})$, alors $\forall y \neq x, \exists \psi, x \succ^{\psi} y$ et $\forall \psi' \rhd \psi, x \succsim^{\psi'} y$. Par l'absurde, supposons que x n'appartient pas à M_K. Il existe donc un plus petit i tel que x n'est pas dans M_i. Par conséquent, il existe $y \in M_i$ tel que $\forall j < i, y \sim^{\Psi_j} x$ (car $x \in M_j$) et $y \succ^{\Psi_i} x$ (car non $x \in M_i$), ce qui contredit l'hypothèse que $x \in M(M_0, \succsim^{\Psi})$. Donc $x \in M_K$.

Réciproquement, soit $x \in M_K$. Par l'absurde, supposons qu'il existe $y \neq x$ tel que $\forall \psi, x \succ^{\psi} y \Rightarrow \exists \psi' \rhd \psi, y \succ^{\psi'} x$. Comme il n'y a pas de point de référence plus prioritaire que ceux de Ψ_1, on a $\forall \psi \in \Psi_1, y \succsim^{\psi} x$. Comme $x \in M_1$, $y \sim^{\Psi_1} x$. Par un raisonnement similaire, on a $\forall \psi \in \Psi_2, y \succsim^{\psi} x$ car sinon on ne pourrait pas trouver de point de référence ψ' plus prioritaire tel que $y \succ^{\psi'} x$. Donc, comme $x \in M_2$, $y \sim^{\Psi_2} x$. Par induction, on montre que $\forall i, y \sim^{\Psi_i} x$, ce qui contredit le fait que $x \neq y$. Par conséquent, $x \in M(M_0, \succsim^{\Psi})$. □

Pour déterminer les solutions préférées, il faudrait calculer itérativement l'ensemble M_K. Toutefois, le calcul direct serait computationnellement inefficace puisque les ensembles M_i sont infinis. Pour que cette opération soit réalisable dans la pratique, nous démontrons une seconde proposition qui indique qu'il suffit de s'intéresser aux politiques stationnaires pures préférées.

Proposition 4.2. *Pour toute solution* $\overline{v}$ *de* M_K*, il existe une politique stationnaire mixte* π *telle que* $\overline{v}^\pi = \overline{v}$*. De plus,* $\overline{v}$ *peut s'exprimer comme une combinaison linéaire de fonctions de valeur de politiques stationnaires pures non dominées au sens de* $\succsim^{\Psi}$.

Démonstration. Les éléments de M_K sont des fonctions de valeur réalisables et de même que pour les MDP classiques, toute fonction de valeur est atteignable par une politique stationnaire mixte.

Soit π une politique stationnaire mixte non dominée pour $\succsim^{\Psi}$. Rappelons que l'ensemble des fonctions de valeur est un ensemble convexe généré par les fonctions de valeur des politiques stationnaires. Comme $\succsim^{\Psi}$ est inclus dans la dominance de Pareto (en considérant tous les points de référence), une solution pour $\succsim^{\Psi}$ est également solution pour la dominance de Pareto. Donc la fonction de valeur de la politique π peut s'écrire comme une combinaison linéaire de fonctions de valeur de politiques stationnaires pures $\delta_1, \ldots, \delta_m$ qui sont toutes Pareto non dominées car dans le cas contraire π serait Pareto-dominée. Formellement, on a

$$\overline{v}^{\pi} = \sum_{i=1}^{m} \lambda_i \overline{v}_i^{\delta} \qquad [15]$$

avec $\sum_{i=1}^{m} \lambda_i = 1$ et $\lambda_i > 0, \forall i = 1, \ldots, m$.

Montrons que chaque δ_i est aussi non dominée pour $\succsim^{\Psi}$. Supposons par l'absurde qu'il existe un δ_j dominé pour $\succsim^{\Psi}$ par une certaine politique π' non dominée pour $\succsim^{\Psi}$. Il existe donc un point de référence ψ tel que $\pi' \succ^{\psi} \delta_j$ et pour tout point de référence $\psi' \rhd \psi$, $\pi' \succsim^{\psi'} \delta_j$. Comme π' est également Pareto non dominée, sa fonction de valeur peut également s'écrire comme combinaison linéaire de fonctions de valeur de politiques stationnaires pures. En remplaçant $\hat{v}^{\delta_i}$ dans l'équation 15 par cette dernière combinaison linéaire, on construit une fonction de valeur $\succsim^{\Psi}$-dominant $\hat{v}^{\pi}$ car les dominances au sens de $\succ^{\psi}$ et $\succsim^{\psi'}$ sont conservées par combinaison linéaire. On obtient donc une contradiction avec le fait que π est non dominée au sens de $\succsim^{\Psi}$. □

D'après cette proposition, il est possible de ne s'intéresser qu'aux politiques stationnaires pures non dominées au sens de $\succsim^{\Psi}$, qui sont en nombre fini. Les autres politiques pourrait être retrouvées à partir de celles-ci.

L'avantage d'utiliser le programme linéaire PL est que l'on peut rajouter aisément de nouvelles contraintes. Pour Ψ, Ψ' deux ensembles de points de référence, $\overline{v} \in \mathbb{R}^k$, appelons le programme linéaire multiobjectif $PL(\Psi, \Psi', \overline{v})$ défini par $PL(\Psi)$ auquel on rajoute une contrainte supplémentaire : $\overline{v}(x) \sim^{\Psi'} \overline{v}$.

Pour déterminer les solutions préférées au sens de $\succsim^{\Psi}$, il suffit donc d'après les propositions précédentes d'appliquer l'algorithme indiqué dans la figure 2.

Supposons qu'on ait résolu $PL(\Psi_1)$, c'est-à-dire le programme linéaire multiobjectif qui correspond à la recherche des politiques non dominées au sens de Pareto quand seuls les points de référence de Ψ_1 sont pris en compte. L'ensemble de ces fonctions de valeur solutions est donc M_1. D'après la proposition 4.2, on peut ne conserver que celles des politiques pures (figure 2, ligne 2) car les fonctions de valeur de toutes les autres politiques (mixtes) préférées s'obtiennent par combinaison linéaire.

Figure 2. *Algorithme de résolution pour le cas : horizon infini, plusieurs points de référence*

1: Résoudre $PL(\Psi_1)$
2: Stocker les fonctions de valeur solutions dans $\overline{V}_1$
3: **for all** $i \in \{2, \ldots, K\}$ **do**
4: Résoudre $PL(\Psi_i, \cup_{k<i}\Psi_k, \overline{v})$ pour chaque $\overline{v} \in \overline{V}_{i-1}$
5: Stocker les fonctions de valeur solutions dans $\overline{V}_i$
6: **end for**
7: Retourner $\overline{V}_K$

Maintenant pour déterminer M_2, il suffit de résoudre $PL(\Psi_2, \Psi_1, \overline{v})$ pour chaque $\overline{v}$ dans M_1. D'après la proposition 4.2, on peut se restreindre à résoudre $PL(\Psi_2, \Psi_1, \overline{v})$ pour les politiques pures dont les fonctions de valeur $\overline{v}$ sont dans M_1 (figure 2, ligne 5). Ensuite il suffit d'itérer ce procédé jusqu'à M_K en ne s'intéressant qu'aux politiques pures.

5. Conclusion

Nous avons étendu le modèle des processus décisionnels de Markov (MDP) pour la prise en compte de récompenses ordinales. Ce modèle est utile dans les situations où les récompenses sont difficiles à évaluer mais également pour celles où la nature des récompenses est réellement qualitative.

À cette fin, nous avons identifié un MDP avec récompenses ordinales à un MDP avec récompenses vectorielles. Dans un tel MDP vectoriel, les fonctions de valeur peuvent se transformer en distributions de probabilité sur les récompenses ordinales. Grâce à cette observation, il est alors naturel d'exploiter des relations de préférence sur les distributions de probabilité dans les MDP vectoriels. Après avoir passé en revue quelques relations de préférence, nous nous sommes intéressé aux dominances probabilistes à points de référence. Avec ces relations, il est possible d'exprimer des préférences assez riches. Nous avons alors proposé une nouvelle méthode de résolution pour la dominance probabiliste avec plusieurs points de référence pour l'horizon infini dans le cas où la priorité sur les points de référence est un préordre total.

Comme extension à ce travail, il serait intéressant de généraliser notre méthode de résolution dans le cas général au cas où la priorité sur les points de référence est simplement un ordre strict. Par ailleurs, d'autres relations de préférence sur les distributions de probabilité pourraient être considérées, notamment celles proposées dans (Perny *et al.,* 2006, Bouyssou *et al.,* 2007).

6. Bibliographie

Bagnell J., Ng A., Schneider J., Solving Uncertain Markov Decision Processes, Technical report, CMU, 2001.

Bouyssou D., Marchant T., « An axiomatic approach to noncompensatory sorting methods in MCDM, II : More than two categories », *EJOR*, vol. 178, p. 246-276, 2007.

Bouyssou D., Marchant T., Perny P., Pirlot M., Tsoukiàs A., Vincke P., *Evaluation and decision models : a critical perspective*, Kluwer, 2000.

Castagnoli E., Li Calzi M., « Expected utility without utility », *Theory and Decision*, vol. 41, p. 281-301, 1996.

Givan R., Leach S., Dean T., « Bounded-parameter Markov decision process », *Artif. Intell.*, vol. 122, n° 1-2, p. 71-109, 2000.

Grosof B., « Generalizing prioritization », *KR*, vol. 2, p. 289-300, 1991.

Junker U., « Preference-Based Search and Multi-Criteria Optimization », *AAAI*, vol. 18, p. 34-40, 2002.

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

Nilim A., El Ghaoui L., « Robustness in Markov Decision Problems with Uncertain Transition Matrices », *NIPS*, 2003.

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

Perny P., Pomerol J.C., « Use of artificial intelligence in MCDM », *in*, T. Gal, T. Stewart, T. Hanne (eds), *Multicriteria Decision Making Advances in MCDM Models, Algorithms Theory, and Applications*, Kluwer Academic, p. 15 :1-15 :43, 1999.

Perny P., Rolland A., « Reference dependent qualitative models for decision making under uncertainty », *ECAI*, p. 422-426, 2006.

Perny P., Spanjaard O., Weng P., « Algebraic Markov Decision Processes », *IJCAI*, vol. 19, p. 1372-1377, 2005.

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.

Shaked M., Shanthikumar J., *Stochastic Orders and Their Applications (Probability and Mathematical Statistics)*, Academic press, 1994.

Sigaud, O., Buffet, O. (eds), *Processus décisionnels de Markov en intelligence artificielle*, Hermès, 2008.

Steuer R., *Multiple criteria optimization*, John Wiley, 1986.

Trevizan F., Cozman F., de Barros L., « Planning under Risk and Knightian Uncertainty », *IJCAI*, p. 2023-2028, 2007.

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

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

Zeleny M., *Linear multiobjective programming*, Springer-Verlag, 1974.