# Processus décisionnels de Markov possibilités à observabilité mixte

Nicolas Drougard, Florent Teichteil-Konigsbuch, Jean-Loup Farges, Didier Dubois

▶ **To cite this version:**

Nicolas Drougard, Florent Teichteil-Konigsbuch, Jean-Loup Farges, Didier Dubois. Processus décisionnels de Markov possibilités à observabilité mixte. Revue des Sciences et Technologies de l'Information - Série RIA : Revue d'Intelligence Artificielle, 2015, vol. 29 (n° 6), pp. 629-653. 10.3166/RIA.29.629-653 . hal-01530407

**HAL Id: hal-01530407**
**https://hal.science/hal-01530407**

Submitted on 31 May 2017

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.


Open Archive TOULOUSE Archive Ouverte (OATAO)
OATAO is an open access repository that collects the work of Toulouse researchers and makes it freely available over the web where possible.

This is an author-deposited version published in : http://oatao.univ-toulouse.fr/
Eprints ID : 16980

**To link to this article** : DOI : 10.3166/RIA.29.629-653
URL : http://dx.doi.org/10.3166/RIA.29.629-653

**To cite this version** : Drougard, Nicolas and Teichteil-Konigsbuch, Florent and Farges, Jean-Loup and Dubois, Didier *Processus décisionnels de Markov possibilités à observabilité mixte.* (2015) Revue d'Intelligence Artificielle, vol. 29 (n° 6). pp. 629-653. ISSN 0992-499X

Any correspondence concerning this service should be sent to the repository administrator: staff-oatao@listes-diff.inp-toulouse.fr

# Processus décisionnels de Markov possibilistes à observabilité mixte

**Nicolas Drougard** [1], **Florent Teichteil-Königsbuch** [1], **Jean-Loup Farges** [1], **Didier Dubois** [2]

1. *Onera – The French Aerospace Lab*
   *2 avenue Edouard Belin*
   *31055 Toulouse, France*
   *nom.prenom@onera.fr*

2. *IRIT – Université Paul Sabatier*
   *118 route de Narbonne*
   *31062 Toulouse, France*
   *dubois@irit.fr*

RÉSUMÉ. *Les processus décisionnels de Markov partiellement observables possibilistes qualitatifs ($\pi$-PDMPO) constituent une alternative aux PDMPO classiques (probabilistes) : ils sont utilisés dans les situations où l'état de croyance initial de l'agent et les probabilités définissant le problème sont imprécises du fait d'un manque de connaissance ou de données. Cependant, tout comme les PDMPO, le calcul d'une stratégie optimale demande un grand nombre d'opérations : le nombre d'états de croyance grandit exponentiellement avec le nombre d'états du système. Dans cet article, une version possibiliste des processus décisisionnels de Markov à observabilité mixte est présentée pour simplifier ce calcul : la complexité de résolution d'un $\pi$-PDMPO, dont certaines variables d'état sont complètement observables, peut être considérablement réduite. Un algorithme d'itération sur les revenus optimaux pour cette nouvelle formulation est ensuite proposé pour le cas de l'horizon infini, et l'optimalité de la stratégie calculée pour un critère donné est démontrée, lorsqu'il existe une action "rester" dans certains états buts. Les expérimentations montrent finalement que ce modèle possibiliste est plus performant que le modèle PDMPO probabiliste, utilisé classiquement en robotique, pour un problème de reconnaissance de cible, dans certaines situations où les capacités d'observation de l'agent ne sont pas précises.*

ABSTRACT. *Possibilistic and qualitative Partially Observable Markov Decision Processes ($\pi$-POMDPs) are counterparts of POMDPs used to model situations where the agent's initial belief and the probabilities defining the problem are imprecise due to lack of past experiences or insufficient data collection. However, like probabilistic POMDPs, optimally solving $\pi$-POMDPs is intractable: the finite belief state space grows exponentially with the number of system states. In this paper, a possibilistic version of Mixed-Observable MDPs is presented to get around this issue: the complexity of solving $\pi$-POMDPs, some state variables of which are fully observable,*

*can be then dramatically reduced. A value iteration algorithm for this new formulation under infinite horizon is next proposed, and the optimality of the returned policy for a specified criterion is shown assuming the existence of a "stay" action in some goal states. Experimental work finally shows that this possibilistic model outperforms probabilistic POMDPs commonly used in robotics, for a target recognition problem where the agent's observation capacities are imprecise.*

MOTS-CLÉS : *PDMPO, observabilité mixte, théorie des possibilités, paramètres imprécis*

KEYWORDS: *POMDPs, Mixed-Observability, Possibility Theory, Imprecise Parameters*

## 1. Introduction

Les processus décisionnels de Markov (PDM) constituent un formalisme adapté aux problèmes de décision séquentielle sous incertitude (Bellman, 1957). Les PDM partiellement observables (PDMPO) (Smallwood, Sondik, 1973) permettent la modélisation des situations dans lesquelles l'agent n'a pas une connaissance directe de l'état courant du système : ses décisions sont alors fondées sur une estimation de cet état, prenant la forme d'une distribution de probabilité définie sur l'espace des états du système. Cette distribution, appelée état de croyance, est mise à jour à chaque étape $t \in \mathbb{N}$ du processus en utilisant l'observation courante du système. Cette mise à jour, effectuée à l'aide de la règle de Bayes, nécessite une connaissance parfaite de l'état de croyance initial de l'agent, ainsi que des distributions de probabilité de transition et d'observation du processus.

Considérons la situation dans laquelle l'agent ignore complètement l'état initial du système. Cette situation s'illustre simplement avec le cas concret d'un robot se trouvant pour la première fois dans une des pièces d'un bâtiment. Son but consiste à atteindre la sortie de la pièce mais il ne connaît pas la position de cette sortie. Son état croyance porte sur cette position. En pratique, aucune expérience ne peut être répétée afin d'extraire une fréquence de position pour la sortie. Dans ce type de situation, l'incertitude n'est pas due à un fait aléatoire, mais à un manque de connaissance : aucun état de croyance initial fréquentiste ne peut permettre la définition du modèle. Il est courant de choisir une probabilité uniforme afin d'attribuer la même probabilité à chaque état, c'est-à-dire à chaque position possible pour la sortie. Ce choix se justifie à travers la théorie des probabilités subjectives (De Finetti, 1937; Dubois *et al.*, 1996) : la probabilité d'une position représente alors le prix $\in [0, 1]$ qu'une personne est prête à donner pour le pari de gagner 1 si c'est la vraie position. Cependant, cette personne prend en compte, dans son choix de prix, que le pari est échangeable. C'est-à-dire qu'elle doit accepter le cas échéant de revendre le pari au même prix que celui auquel elle propose de l'acheter. Cette définition des probabilités subjectives à partir de ces paris échangeables est différente de la définition fréquentiste des probabilités. Le choix d'un état de croyance initial uniforme force alors sa mise à jour à mélanger

des probabilités initiales subjectives avec des fréquences d'observations, ce qui n'a pas toujours de sens.

Dans une autre situation, l'agent pourrait croire fortement que la sortie de la pièce est située dans un mur, comme c'est le cas pour la plupart des pièces. Il conserverait cependant une toute petite probabilité $p_\epsilon$ associée au fait que la sortie soit une trappe au milieu de la pièce. Même si cette situation est très peu probable, cette seconde option doit être prise en compte dans l'état de croyance. Dans le cas contraire, la règle de Bayes ne pourrait pas mettre à jour cet état de croyance correctement, si la sortie est effectivement au milieu de la pièce. Il n'est pas du tout évident de définir $p_\epsilon$ sans expérience passée : cela ne se baserait sur aucune justification rationnelle. Cependant, la stratégie de l'agent dépendra de cette définition. Contrairement aux probabilités, les modèles d'incertitude possibilistes permettent la modélisation des états de croyance avec connaissance imprécise.

De plus, dans le cas d'une mission robotique utilisant la perception visuelle, les observations du robot sont les sorties d'un algorithme de traitement d'image dont la mécanique peut inclure une corrélation entre images, une mise en correspondance avec des objets, un apprentissage basé sur un ensemble d'images représentatives, ou non, des situations rencontrées à l'exécution,... Cette mécanique est tellement complexe, que des probabilités d'occurrence seraient difficiles à extraire rigoureusement.

Le modèle $\pi$-PDMPO (Sabbadin, 1999) est une alternative possibiliste et qualitative du modèle probabiliste PDMPO classique : il permet une modélisation formelle de l'ignorance totale, en utilisant une distribution de possibilité égale à 1 pour tous les états. Cette distribution signifie que les états sont tous également possibles, mais qu'aucun n'est nécessaire. De plus, extraire des estimations qualitatives des performances de reconnaissance d'un algorithme de traitement d'image est plus facile : le modèle $\pi$-PDMPO ne nécessite qu'un paramétrage qualitatif, donc il permet de construire le modèle sans utiliser plus d'informations que celles vraiment disponibles.

Cependant, tout comme le modèle probabiliste PDMPO, ce modèle possibiliste est très complexe à résoudre, c'est-à-dire le calcul d'une stratégie optimale peut s'avérer impossible en pratique. En effet, la taille de l'espace des états de croyances grandit exponentiellement avec la taille de l'espace d'état, ce qui empêche l'utilisation des $\pi$-PDMPO en pratique. Dans les situations où une large partie de l'état est complètement observable, un agencement différent du modèle permet de résoudre plus facilement le problème : le modèle possibiliste PDM à observabilité mixte ($\pi$-PDMOM), qui est la première contribution de cet article, permet un raisonnement avec des états de croyance sur les états partiellement observés seulement. Dans ce modèle, emprunté aux PDMOM probabilistes (Ong *et al.*, 2010; Araya-Lòpez *et al.*, 2010), les états sont factorisés en une variables d'état visible, et une variable d'état partiellement observable, ou cachée. Pour les PDMPO probabilistes, ce modèle factorisé permet de raisonner sur des plus petits sous-espaces *continus* d'états de croyance et accélère les opérations sur les $\alpha$-vecteurs. L'impact est complètement différent pour les PDMPO possibilistes qualitatifs : la factorisation permet de réduire la taille de l'espace *fini* des états de croyance.

La seconde contribution consiste en un algorithme d'itération sur les revenus optimaux pour cette extension, qui exploite la structure hybride de l'espace des états de croyance, induite par la factorisation de la variable visible avec la variable cachée. Cet algorithme est dérivé de l'algorithme pour les $\pi$-PDM, mettant à jour le travail de Sabbadin, dont l'optimalité de la stratégie calculée est prouvée pour un critère d'horizon infini explicité. Une action "rester" intermédiaire est nécessaire pour garantir la convergence de l'algorithme, mais elle disparaît dans la stratégie optimale pour les états qui ne sont pas des buts ; cette action est l'équivalent possibiliste du coefficient d'actualisation des PDMPO probabilistes.

Enfin, les expérimentations visent à démontrer l'hypothèse que dans certaines situations, les $\pi$-PDMOM peuvent être plus performants que leur équivalents probabilistes, par exemple dans des problèmes robotiques de collecte d'informations, lorsque la fonction d'observation n'est pas connue précisément, comme c'est souvent le cas dans les applications réelles. La validation de cette hypothèse est significative car il est communément admis dans le domaine de la robotique que les PDMPO probabilistes, et plus généralement les approches bayésiennes, sont des modèles à privilégier pour résoudre des missions de collecte d'information. Nous apportons dans cet article des éléments indiquant que, en pratique, les modèles d'incertitude possibilistes peuvent mieux se comporter que leurs équivalents probabilistes.

## 2. Contexte

Les processus décisionnels de Markov modélisent les situations dans lesquelles un *système*, par exemple la partie physique d'un agent et son environnement, a une dynamique markovienne dans le temps. Les différents états possibles du système sont représentés par les éléments $s$ de l'espace fini des états $\mathcal{S}$. L'état initial du système est noté $s_0 \in \mathcal{S}$. A chaque étape du processus, modélisé par les entiers $t \in \mathbb{N}$, la partie décisionnelle de l'agent peut choisir une *action* notée $a$ dans l'ensemble fini $\mathcal{A}$. L'action choisie au temps $t$, $a_t$, détermine l'incertitude sur l'état suivant $s_{t+1}$ connaissant l'état courant $s_t$.

Dans le cadre probabiliste cette incertitude est décrite par une fonction de transition donnant la probabilité de l'état suivant $s_{t+1}$ conditionnellement à l'état courant $s_t$ et à l'action $a_t$, notée $\mathbf{p}(\, s_{t+1} \mid\ s_t, a_t\,)$. La sélection des actions se base sur un critère amenant le système à parcourir une suite d'états satisfaisants. Ce critère utilise la notion de récompense : la fonction de récompense $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ associe à chaque couple état-action $(s, a) \in \mathcal{S} \times \mathcal{A}$ la valeur $r(s, a)$, appelée récompense. Une suite d'actions est alors dite optimale si la somme (ou une agrégation semblable) des récompenses est la plus grande en moyenne : le critère est donc l'espérance de cette somme. Quand l'état du système est partiellement observable, c'est-à-dire lorsque le problème est un PDMPO (Smallwood, Sondik, 1973), l'agent n'a plus accès aux états du système durant l'exécution du processus, mais seulement à des observations $o_{t+1}$ qui permettent de l'estimer, à l'aide de la fonction d'observation $\mathbf{p}(\, o_{t+1} \mid\ s_{t+1}, a_t\,)$. Cependant, le critère d'optimisation reste le même.

Dans le cas des PDM et PDMPO possibilistes qualitatifs décrits dans la suite, les fonctions de transition et d'observation ne sont plus probabilistes (et donc quantitatives), mais qualitatives, tout comme la fonction de récompense, appelée alors *préférence*. Supposons que les probabilités régissant un système robotique ne sont pas connues, mais qu'un expert de ce système puisse classer tous les événements possibles par ordre de plausibilité: les informations concernant le système sont alors des données qualitatives, et même des *distributions de possibilité qualitatives*.

### *2.1. PDMs possibilistes qualitatifs*

Le travail de Sabbadin (1999) propose un homologue possibiliste aux PDM. Dans ce cadre, l'incertitude concernant les transitions est modélisée par des distributions de possibilité qualitatives sur $\mathcal{S}$. Soit $\mathcal{L}$ l'échelle possibiliste, c'est-à-dire un ensemble fini et totalement ordonné dont le plus grand élément est noté $1_{\mathcal{L}}$ et le plus petit $0_{\mathcal{L}}$. Un exemple simple est $\mathcal{L} = \left\{ 0, \frac{1}{k}, \frac{2}{k}, \ldots, \frac{k-1}{k}, 1 \right\}$ avec $k \in \mathbb{N}^*$.

Une distribution de possibilité qualitative sur $\mathcal{S}$ est une fonction $\pi : \mathcal{S} \to \mathcal{L}$ telle que $\max_{s \in \mathcal{S}} \pi(s) = 1_{\mathcal{L}}$. Cette normalisation possibiliste implique qu'au moins un état $s$ est entièrement possible. L'inégalité $\pi(\overline{s}) < \pi(\tilde{s})$ signifie que l'état $\tilde{s}$ est plus plausible que l'état $\overline{s}$. Cette modélisation nécessite moins d'information que la modélisation probabiliste, puisque les plausibilités des événements sont seulement classées dans $\mathcal{L}$ et non pas quantifiées.

La fonction de transition $T^\pi$ est définie comme suit : pour chaque paire d'états $(s, s') \in \mathcal{S}^2$ et pour l'action $a \in \mathcal{A}$, $T^\pi(s, a, s') = \pi\,(\,s' \mid\ s, a\,) \in \mathcal{L}$ est le degré de possibilité que l'état du système devienne $s'$ conditionnellement à l'état courant $s$ et à l'action $a$. L'échelle $\mathcal{L}$ sert aussi à modéliser la *préférence* sur les états : la fonction de préférence $\mu : \mathcal{S} \to \mathcal{L}$ modélise le but de la mission de l'agent, c'est-à-dire les états dans lesquels l'agent doit amener le système. Un $\pi$-PDM est entièrement défini avec le 5-uplet $\langle \mathcal{S}, \mathcal{A}, \mathcal{L}, T^\pi, \mu \rangle$.

Une *stratégie* est une suite $(\delta_t)_{t \geqslant 0}$ de *règles de décision* $\delta : \mathcal{S} \to \mathcal{A}$ indexées par l'étape du processus $t \in \mathbb{N}$ : $\delta_t(s)$ est l'action à exécuter dans l'état $s$ au temps $t$ du processus. L'ensemble des stratégies de longueur $p$, $(\delta_0, \ldots, \delta_{p-1})$, est noté $\Delta_p$.

Soit $\tau = (s_1, \ldots, s_p) \in \mathcal{S}^p$ une trajectoire de longueur $p$, et $(\delta) = (\delta_t)_{t=0}^{p-1}$ une stratégie de longueur $p$. L'ensemble de toutes les trajectoires de longueur $p$ est noté $\mathcal{T}_p$.

Pour $(\delta)$ fixé, la suite de variables $(s_t)_{t \geqslant 0}$ est un processus de Markov : le degré de possibilité de la trajectoire $\tau = (s_1, \ldots, s_p)$ qui commence par $s_0$ exécutant la stratégie $(\delta) \in \Delta_p$ est alors

$$\Pi\,(\,\tau \mid\ s_0, (\delta)\,) = \min_{t=0}^{p-1} \pi\,(\,s_{t+1} \mid\ s_t, \delta_t(s_t)\,)\,.$$

La préférence de $\tau \in \mathcal{T}_p$ est définie comme étant la préférence du dernier état : $M(\tau) = \mu(s_p)$. Dans le cas de l'horizon fini, les travaux de Sabbadin *et al.* (1998)

et de Sabbadin (1999) proposent de définir la préférence d'une trajectoire comme le minimum des préférences de chaque état. Cependant cette approche mène à un effet de noyade, comme décrit par exemple par Dubois, Fortemps (2005): en effet, deux trajectoires peuvent être considérées comme équivalentes en termes de préférence parce qu'elles ont le même minimum, alors que l'une des deux a des préférences supérieures à celles de l'autre. Cet effet peut être évité à l'aide d'agrégations plus discriminantes. Le cadre qualitatif permet par exemple de définir la préférence d'une trajectoire à l'aide d'opérateurs lexicographiques : *leximin* ou *leximax*. Considérons deux trajectoires de taille $p > 0$, $\tau_1$ et $\tau_2$. Notons $(\mu_1^1, \mu_2^1, \ldots, \mu_p^1)$ et $(\mu_1^2, \mu_2^2, \ldots, \mu_p^2)$ les préférences de chacun de leurs états, de telle sorte que $\forall i \in \{1, 2\}$, $\mu_1^i \leqslant \mu_2^i \leqslant \ldots \leqslant \mu_p^i$. La trajectoire $\tau_1$ sera considérée comme étant préférée à $\tau_2$ selon l'opérateur leximin, si il existe $j \in \{1, \ldots, p\}$ tel que $\forall k \in \{1, \ldots, j-1\}$, $\mu_k^1 = \mu_k^2$ et $\mu_j^1 > \mu_j^2$. Selon l'opérateur leximax, si il existe $j \in \{1, \ldots, p\}$ tel que $\forall k \in \{j+1, \ldots, p\}$, $\mu_k^1 = \mu_k^2$ et $\mu_j^1 > \mu_j^2$, la trajectoire $\tau_1$ sera préférée. La généralisation des PDM possibilistes (Weng, 2007) ainsi que le travail de Dubois, Fortemps (2005) sur l'opérateur leximin pour la décision peuvent alors être utilisés pour formaliser ces préférences pour des problèmes à horizon fini. Cependant, le cas de l'horizon infini est mal connu pour de tels critères : ces opérateurs définissant un ordre total sur les trajectoires, la taille de l'échelle permettant de les classer croît strictement avec l'horizon, et ne permet pas de rester dans le cadre établi précédemment. Il est possible de définir les opérateurs $n$-leximin (respectivement $n$-leximax), comparant uniquement les $n$ plus petites (respectivement plus grandes) préférences de chaque trajectoire afin de garder une échelle $\mathcal{L}$ finie. Le travail présenté ici se concentre en revanche sur le modèle de préférences terminales pour lequel les opérateurs leximin et leximax sont inutiles, et propose une preuve de l'optimalité de la stratégie dans ce cas.

Comme conseillé par Sabbadin (2000), lorsque le système en question ne risque pas d'être bloqué dans des états non satisfaisants, le critère de décision qualitatif optimiste (Dubois, Prade, 1995) est utilisé : l'intégrale de Sugeno de la distribution de préférence sur les trajectoires contre la mesure de possibilité :

$$u_p\left(s_0, (\delta)\right) = \max_{\tau \in \mathcal{T}_p} \min\left\{\Pi\left(\tau \mid s_0, (\delta)\right), M(\tau)\right\}. \tag{1}$$

Une stratégie qui maximise ce critère assure qu'il existe une trajectoire de longueur $p$ possible et satisfaisante. Le $\pi$-PDM à horizon fini est résolu lorsqu'une telle stratégie est calculée. Le critère optimisé pour un horizon $p$, $u_p^*(s) = \max_{(\delta)\in\Delta_p} u_p(s, (\delta))$, est la solution de l'équation de programmation dynamique suivante, qui calcule une stratégie $(\delta^*)$ optimale comme démontré par Sabbadin *et al*. (1998) : $\forall i \in \{1 \ldots p\}$, $\forall s \in \mathcal{S}$,

$$u_i^*(s) = \max_{a \in \mathcal{A}} \max_{s' \in \mathcal{S}} \min\left\{\pi\left(s' \mid s, a\right), u_{i-1}^*(s')\right\}, \tag{2}$$

$$\delta_{p-i}^*(s) \in \operatorname*{argmax}_{a \in \mathcal{A}} \max_{s' \in \mathcal{S}} \min\left\{\pi\left(s' \mid s, a\right), u_{i-1}^*(s')\right\}$$

$$\text{avec l'initialisation } u_0^*(s) = \mu(s).$$

### *2.2. Le cas de l'observation partielle*

Un équivalent possibiliste des PDMPO a été construit par Sabbadin (1999). Comme dans le cadre probabiliste, l'agent n'observe pas directement les états du système et les différentes observations possibles du système sont représentées par les éléments $o$ de l'espace fini des observations $\mathcal{O}$. L'incertitude sur les observations est ici décrite par des distributions de possibilité. La fonction d'observation $\Omega^\pi$ est définie comme suit : $\forall o' \in \mathcal{O}$, $s' \in \mathcal{S}$ et $a \in \mathcal{A}$, $\Omega^\pi(s', a, o') = \pi(o' \mid s', a)$ la possibilité de l'observation courante $o'$ conditionnée par l'état courant $s'$ et l'action précédente $a$. Un $\pi$-PDMPO est alors entièrement défini par le 8-uplet $\langle \mathcal{S}, \mathcal{A}, \mathcal{L}, T^\pi, \mathcal{O}, \Omega^\pi, \mu, \beta_0 \rangle$, où $\beta_0$ est l'état de croyance initial. L'état de croyance de l'agent est une distribution de possibilité sur les états $\mathcal{S}$ ; l'ignorance totale est définie par un état de croyance égal à $1_\mathcal{L}$ pour tous les états, tandis qu'un état donné $s$ est parfaitement connu si l'état de croyance est égal à $1_\mathcal{L}$ pour cet état, et à $0_\mathcal{L}$ pour tous les autres.

La traduction du $\pi$-PDMPO en $\pi$-PDM est similaire à celle utilisée pour les PDMPO classiques : notons $\mathcal{B}^\pi \subset \mathcal{L}^\mathcal{S}$ l'espace des états de croyance possibilistes contenant toutes les distributions de possibilité définies sur $\mathcal{S}$. Les états de croyance possibilistes sont mis à jour comme suit : si au temps $t$, l'état de croyance courant est $\beta_t \in \mathcal{B}^\pi$ et l'agent exécute l'action $a_t$, l'état de croyance sur les états suivants est

$$\beta_{t+1}^{a_t}(s') = \max_{s \in \mathcal{S}} \min \{ \pi(s' \mid s, a_t), \beta_t(s) \},$$

et l'état de croyance sur les observations

$$\beta_{t+1}^{a_t}(o') = \max_{s' \in \mathcal{S}} \min \left\{ \pi(o' \mid s', a_t), \beta_{t+1}^{a_t}(s') \right\}.$$

Ensuite, si l'agent observe $o_{t+1} \in \mathcal{O}$, l'équivalent possibiliste de la règle de Bayes assure que

$$\beta_{t+1}(s') = \begin{cases} 1_\mathcal{L} & \text{si } \beta_{t+1}^{a_t}(o_{t+1}) = \pi(s', o_{t+1} \mid \beta_t, a_t) > 0_\mathcal{L} \\ \pi(s', o_{t+1} \mid \beta_t, a_t) & \text{sinon} \end{cases}, \tag{3}$$

où $\forall (s', o') \in \mathcal{S} \times \mathcal{O}$, $\pi(s', o' \mid \beta, a) = \min \{ \pi(o' \mid s', a), \beta^a(s') \}$ est la distribution de possibilité jointe de $(s', o')$. La mise à jour d'un état de croyance $\beta$ est notée $\beta^{a,o'}$ : $\beta_{t+1} = \beta_t^{a,o'}$. La mise à jour de l'état de croyance étant définie, sa dynamique peut être calculée : soit $\Gamma^{\beta,a}(\beta') = \left\{ o' \in \mathcal{O} \mid \beta^{a,o'} = \beta' \right\}$. Alors, la possibilité de transition de l'état de croyance $\beta$ à $\beta'$ est $\pi(\beta' \mid \beta, a) = \max_{o' \in \Gamma^{\beta,a}(\beta')} \beta^a(o')$ avec la convention $\max_\emptyset = 0_\mathcal{L}$.

La préférence associée à chaque état de croyance est définie sous une forme pessimiste afin de préférer les états de croyance informatifs. Une bonne préférence est associée à un état de croyance, lorsqu'il est peu plausible que le système soit dans un état peu satisfaisant : $\mu(\beta) = \min_{s \in \mathcal{S}} \max \{ \mu(s), n(\beta(s)) \}$, avec $n : \mathcal{L} \to \mathcal{L}$ la fonction inversant l'ordre de $\mathcal{L}$, c'est-à-dire la seule fonction strictement décroissante

de $\mathcal{L}$ dans $\mathcal{L}$. Un $\pi$-PDM sur les état de croyances $\mathcal{B}^\pi$ est alors défini, et la nouvelle équation de programmation dynamique est $\forall i \in \{1, \ldots, p\}, \forall \beta \in \mathcal{B}^\pi$,

$$\begin{aligned} u_p^*(\beta) &= \max_{a \in \mathcal{A}} \max_{\beta' \in \mathcal{B}^\pi} \min\left\{\pi\left(\beta' \mid \beta, a\right), u_{p-1}^*(\beta')\right\} \\ &= \max_{a \in \mathcal{A}} \max_{o' \in \mathcal{O}} \min\left(\beta^a(o'), u_{p-1}^*(\beta^{a,o'})\right), \end{aligned}$$

avec l'initialisation $u_0^*(\beta) = \mu(\beta)$. Notons que $\mathcal{B}^\pi$ est un ensemble fini de taille $\#\mathcal{B}^\pi = \#\mathcal{L}^{\#\mathcal{S}} - (\#\mathcal{L}-1)^{\#\mathcal{S}}$ : le nombre total de vecteurs de taille $\#\mathcal{S}$ à valeurs dans $\mathcal{L}$, moins $(\#\mathcal{L}-1)^{\#\mathcal{S}}$ distributions non normalisées. Cependant, pour des problèmes concrets, l'espace d'état peut être très grand : $\#\mathcal{B}^\pi$ explose et les calculs deviennent insurmontables, comme pour les PDMPO probabilistes. La prochaine section présente la première contribution de ce papier, qui exploite une structure particulière du problème qui est très fréquente en pratique.

## 3. PDM possibilistes qualitatifs à observabilité mixte ($\pi$-PDMOM)

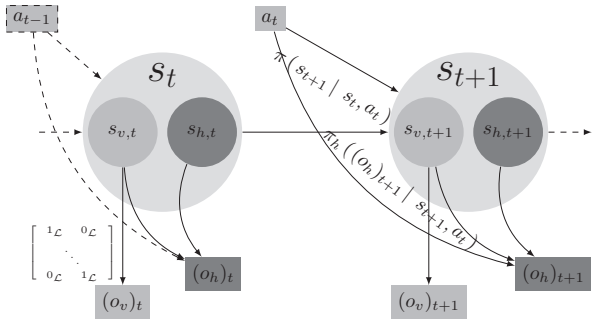



*Figure 1. Représentation graphique d'un $\pi$-PDMOM*

La complexité de la résolution des $\pi$-PDMPO est due au fait que la taille de l'espace des états de croyance $\#\mathcal{B}^\pi$ grandit exponentiellement avec la taille de l'espace des états $\#\mathcal{S}$. Cependant, en pratique, les états sont rarement totalement cachés. L'exploitation de l'observabilité mixte peut être une solution : inspirée par un travail récent à propos des PDMPO probabilistes de Ong *et al*. (2010) et de Araya-Lòpez *et al*. (2010), cette section présente une modélisation structurée prenant en compte les situations où l'agent observe directement certaines parties de l'état du système. Ce modèle généralise alors les $\pi$-PDM et les $\pi$-PDMPO.

Comme dans les travaux de Ong *et al*. (2010) et Araya-Lòpez *et al*. (2010), nous faisons l'hypothèse que l'espace d'état $\mathcal{S}$ peut être écrit comme le produit cartésien d'un espace d'états visibles $\mathcal{S}_v$ et d'un espace d'états cachés $\mathcal{S}_h$ : $\mathcal{S} = \mathcal{S}_v \times \mathcal{S}_h$. Soit $s = (s_v, s_h)$ un état du système. La composante $s_v \in \mathcal{S}_v$ est visible par l'agent et $s_h \in \mathcal{S}_h$ est seulement partiellement observée à travers les observations de l'ensemble $\mathcal{O}_h$ : la distribution de possibilité sur la future observation $o'_h \in \mathcal{O}_h$ sachant l'état

futur $s' \in \mathcal{S}$ et l'action courante $a \in \mathcal{A}$ est notée $\pi\,(\,o'_h \mid\ s', a\,)$. La figure 1 illustre ce modèle structuré.

Vu comme un $\pi$-PDMPO, l'espace des états visibles est intégré à l'espace des observations : $\mathcal{O}_v = \mathcal{S}_v$ et $\mathcal{O} = \mathcal{O}_v \times \mathcal{O}_h$. Ainsi, sachant que l'état courant du système est $s'_v$, l'agent observe *nécessairement* $o'_v = s'_v$. C'est à dire que si $o'_v \neq s'_v$, $\pi\,(\,o'_v \mid\ s'_v\,) = 0_{\mathcal{L}}$. Formellement, dans le cadre des $\pi$-PDMPO, la distribution de possibilité sur les observations s'écrit :

$$\begin{aligned} \pi\,(\,o' \mid\ s', a\,) &= \pi\,(\,o'_v, o'_h \mid\ s'_v, s'_h, a\,) \\ &= \min\{\,\pi\,(\,o'_h \mid\ s'_v, s'_h, a\,), \pi\,(\,o'_v \mid\ s'_v\,)\,\} \\ &= \begin{cases} \pi\,(\,o'_h \mid\ s', a\,) & \text{si } o'_v = s'_v, \\ 0_{\mathcal{L}} & \text{sinon,} \end{cases} \end{aligned} \tag{4}$$

puisque $\pi\,(\,o'_v \mid\ s'_v\,) = 1_{\mathcal{L}}$ si $s'_v = o'_v$ et $0_{\mathcal{L}}$ sinon. Le théorème suivant, basé sur cette égalité, permet de définir l'état de croyance de l'agent sur les états cachés du système.

THÉORÈME 1 (Réécriture de l'état de croyance). — *Tout état de croyance atteignable d'un $\pi$-PDMOM peut s'écrire comme un élément de $\mathcal{S}_v \times \mathcal{B}^\pi_h$ où $\mathcal{B}^\pi_h$ est l'ensemble des distributions de possibilité sur $\mathcal{S}_h$ : tout $\beta \in \mathcal{B}^\pi$ peut s'écrire $(s_v, \beta_h)$ avec $\beta_h(s_h) = \max_{\overline{s}_v \in \mathcal{S}_v} \beta(\overline{s}_v, s_h)$ et $s_v = \operatorname{argmax}_{\overline{s}_v \in \mathcal{S}_v} \beta(\overline{s}_v, s_h)$.*

PREUVE. — Raisonnons par récurrence sur $t \in \mathbb{N}$ : puisque l'état visible initial $s_{v,0}$ est connu par l'agent, seuls les états $s = (s_v, s_h)$ pour lesquels $s_v = s_{v,0}$ sont tels que $\beta_0(s) > 0_{\mathcal{L}}$. Un état de croyance sur les états cachés peut alors être défini par $\beta_{h,0}(s_h) = \max_{s_v \in \mathcal{S}_v} \beta_0(s_v, s_h) = \beta_0(s_{v,0}, s_h)$.

À l'étape de temps $t \in \mathbb{N}$, si $\beta_t(s) = 0_{\mathcal{L}}$ pour chaque $s = (s_v, s_h) \in \mathcal{S}$ tel que $s_v \neq s_{v,t}$, la même notation peut être adoptée : $\beta_{h,t}(s_h) = \beta_t(s_{v,t}, s_h)$. Donc, si le système atteint l'état $s_{t+1} = (s_{v,t+1}, s_{h,t+1})$ et si $s' = (s'_v, s'_h)$ avec $s'_v \neq s_{v,t+1}$, alors $s'_v \neq o_{v,t+1}$ et $\pi\,(\,o_{t+1}, s' \mid\ \beta_t, a_t\,) = \min\left\{\,\pi\,(\,o_{t+1} \mid\ s', a_t\,), \beta^{a_t}_{t+1}(s')\,\right\} = 0_{\mathcal{L}}$ grâce à l'équation (4). Enfin, la formule de mise à jour (3) assure que $\beta_{t+1}(s') = 0_{\mathcal{L}}$. Ainsi, $\beta_{t+1}$ est entièrement spécifiée par $(s_{v,t+1}, \beta_{h,t+1})$ avec $s_{v,t+1} = o_{v,t+1}$ et $\beta_{h,t+1}(s_h) = \max_{s_v} \beta_{t+1}(s_v, s_h)\ \forall s_h \in \mathcal{S}_h$. L'état visible suivant $s_{v,t+1}$ est aussi le seul élément de $\mathcal{S}_v$ tel que il existe $s_h \in \mathcal{S}_h$, pour lequel $\beta_{t+1}(s_{v,t+1}, s_h) > 0_{\mathcal{L}}$. ■

Notons néanmoins que les états de croyance sur les états suivants considérés dans les calculs intermédiaires $\beta^{a_t}_{t+1}(s')$ ne sont en général pas concernés par ce théorème. Comme tous les états de croyance atteignables sont en fait dans $\mathcal{S}_v \times \mathcal{B}^\pi_h$, le théorème suivant redéfinit l'équation de programmation dynamique en la restreignant à cet espace produit.

THÉORÈME 2 (nouvelle équation de programmation dynamique). — *Sur $\mathcal{S}_v \times \mathcal{B}^\pi_h$, l'équation de programmation dynamique est initialisée comme suit, $u^*_0(s_v, \beta_h) = \mu(s_v, \beta_h)$, et devient : $\forall i \in \{1, \ldots, p\}$,*

$$u^*_i(s_v, \beta_h) = \max_{a \in \mathcal{A}} \max_{s'_v \in S_v} \max_{o'_h \in O_h} \min\left\{\,\beta^a(s'_v, o'_h), u^*_{i-1}(s'_v, \beta^{a, s'_v, o'_h}_h)\,\right\}$$

*où* $\mu(s_v, \beta_h) = \min_{s_h \in \mathcal{S}_h} \max \{ \mu(s_v, s_h), n(\beta_h(s_h)) \}$ *est la préférence sur* $S_v \times \mathcal{B}_h^\pi$, *et où* $\beta^a(s'_v, o'_h) = \max_{s'_h \in \mathcal{S}_h} \min \{ \pi(o'_h \mid s'_v, s'_h, a), \beta^a(s'_v, s'_h) \}$ *est l'état de croyance sur les variables visibles, c'est-à-dire l'état visible* $s_v$ *et l'observation* $o_h$. *La mise à jour de l'état de croyance devient alors*

$$\beta_h^{s'_v, o'_h, a}(s'_h) = \begin{cases} 1_{\mathcal{L}} & \textit{si } \min \{ \pi(o'_h \mid s'_v, s'_h, a), \beta^a(s'_v, s'_h) \} = \beta^a(s'_v, o'_h) > 0_{\mathcal{L}} \\ \min\{\pi(o'_h \mid s'_v, s'_h, a), \beta^a(s'_v, s'_h)\} & \textit{sinon.} \end{cases}$$

PREUVE. — En utilisant l'équation de programmation dynamique classique, le théorème 1, et le fait que $\mathcal{S}_v = \mathcal{O}_v$,

$$\begin{aligned} u_i^*(s_v, \beta_h) = u_i^*(\beta) &= \max_{a \in \mathcal{A}} \max_{(o'_v, o'_h) \in \mathcal{O}} \min \left\{ \beta^a(o'_v, o'_h), u_{i-1}^*(\beta^{a,(o'_v,o'_h)}) \right\} \\ &= \max_{a \in \mathcal{A}} \max_{s'_v \in \mathcal{S}_v} \max_{o'_h \in \mathcal{O}_h} \min \left\{ \beta^a(s'_v, o'_h), u_{i-1}^*(\beta^{a,s'_v,o'_h}) \right\} \\ &= \max_{a \in \mathcal{A}} \max_{s'_v \in \mathcal{S}_v} \max_{o'_h \in \mathcal{O}_h} \min \left\{ \beta^a(s'_v, o'_h), u_{i-1}^*(s'_v, \beta_h^{a,s'_v,o'_h}) \right\} \end{aligned}$$

où $\forall s_h \in \mathcal{S}_h$, $\beta_h^{a,s'_v,o'_h}(s_h) = \max_{\overline{s}_v} \beta^{a,s'_v,o'_h}(\overline{s}_v, s_h) = \beta^{a,s'_v,o'_h}(s'_v, s_h)$. Pour l'initialisation, notons juste que $n(\beta(\overline{s}_v, s_h)) = 1_{\mathcal{L}}$ lorsque $\overline{s}_v \neq s_v$, et donc $\mu(\beta) = \min_s \max \{ \mu(s), n(\beta(s)) \} = \min_{s_h} \max \{ \mu(s_v, s_h), n(\beta(s_v, s_h)) \}$ puisqu'un minimum différent de $1_{\mathcal{L}}$ ne peut être atteint que pour $\overline{s}_v = s_v$. La nouvelle formule pour la préférence sur $\mathcal{S}_v \times \mathcal{B}_h^\pi$ est ainsi expliquée. L'état de croyance sur les observations définie dans la dernière section peut s'écrire : $\forall o' = (o'_v, o'_h) \in \mathcal{O}$,

$$\begin{aligned} \beta^a(o') &= \max_{s' \in \mathcal{S}} \min \{ \pi(o'_v, o'_h \mid s', a), \beta_t^{a_t}(s') \} \\ &= \max_{s'_h \in \mathcal{S}_h} \min \{ \pi(o'_h \mid s'_v, s'_h, a_t), \beta_t^{a_t}(s'_v, s'_h) \} \end{aligned}$$

avec $s'_v = o'_v$ puisque, dans le cas contraire, $\pi(o' \mid s'_v, s'_h, a) = 0_{\mathcal{L}}$ selon l'équation (4). Ainsi $\beta^a(s'_v, o'_h) = \beta^a(o'_v, o'_h)$. Finalement, en utilisant l'équation de mise à jour standard (3) avec $o'_v = s'_v$ et l'équation (4), nous obtenons la nouvelle mise à jour de l'état de croyance. ■

Un algorithme standard aurait calculé $u_p^*(\beta)$ pour chaque état de croyance $\beta \in \mathcal{B}^\pi$ tandis que cette nouvelle équation de programmation dynamique mène à un algorithme qui calcule ce critère optimal seulement pour les éléments de $\mathcal{S}_v \times \mathcal{B}_h^\pi$. La taille de ce nouvel espace d'états est $\#(\mathcal{S}_v \times \mathcal{B}_h^\pi) = \#\mathcal{S}_v \cdot \left( \#\mathcal{L}^{\#\mathcal{S}_h} - (\#\mathcal{L} - 1)^{\#\mathcal{S}_h} \right)$, ce qui est exponentiellement plus petit que la taille de l'espace des états de croyances du $\pi$-PDMPO associé : $\#\mathcal{B}^\pi = \#\mathcal{L}^{\#\mathcal{S}_v \cdot \#\mathcal{S}_h} - (\#\mathcal{L} - 1)^{\#\mathcal{S}_v \cdot \#\mathcal{S}_h}$.

## 4. Résoudre les $\pi$-PDMOM

Une stratégie pour un nombre fini d'étapes peut maintenant être calculée pour des problèmes à observabilité mixte dont l'espace d'état est plus grand, en utilisant

l'équation de programmation dynamique du théorème 2 et en sélectionnant les actions maximisantes pour chaque état $(s_v, \beta_h) \in \mathcal{S}_v \times \mathcal{B}_h^\pi$, comme décrit par l'équation (2) pour chaque $s \in \mathcal{S}$. Cependant pour de nombreux problèmes en pratique, il est difficile de déterminer l'horizon fini du problème c'est-à-dire le nombre $p$ avant la fin du processus. Le but de cette section est de présenter un algorithme pour résoudre les $\pi$-PDMOM avec un horizon infini : cet algorithme est le premier qui résout, preuve à l'appui, les $\pi$-PDM(OM).

### *4.1. Le cas des $\pi$-PDM*

Un travail précédent sur la résolution des $\pi$-PDM, (Sabbadin, 1999, 2001), a proposé un algorithme d'itération sur les revenus optimaux : il est prouvé que cet algorithme calcule le Revenu optimal, mais pas nécessairement des stratégies optimales pour certains problèmes avec des cycles. Il y a un problème similaire avec les PDM probabilistes sans facteur d'actualisation $\gamma$. Le facteur d'actualisation est un réel dans $[0, 1]$ qui permet de faire converger la somme totale des récompenses $\sum_{t \geqslant 0} \gamma^t \cdot r(s, a)$ dont l'espérance constitue le critère (fonction valeur) du PDM. Ce facteur d'actualisation permet aussi de faire converger l'algorithme d'itération sur les revenus. Sans facteur d'actualisation, la stratégie gourmande après convergence de l'itération sur les revenus optimaux n'est pas forcément optimale (Puterman, 1994). Il n'est pas surprenant que nous rencontrions le même problème avec les $\pi$-PDM Comme, de plus, l'opérateur de programmation dynamique procède à des calculs purement qualitatifs, il ne peut pas être rendu contractant par un facteur d'actualisation $0 < \gamma < 1$.

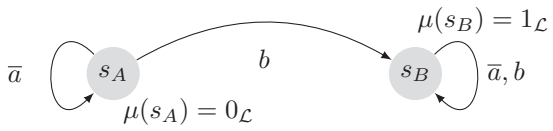



*Figure 2. Exemple de $\pi$-PDM pour lequel toutes les actions sont gourmandes*

A notre connaissance, nous proposons ici le premier algorithme d'itération sur les revenus optimaux pour les $\pi$-PDM, dont il est prouvé qu'il renvoie une stratégie optimale, et qui est différent de celui présenté par Sabbadin (2001). En effet, dans l'exemple déterministe de la figure 2, l'action $\overline{a}$, qui est clairement sous optimale, est déclarée optimale dans l'état $s_A$ par cet algorithme: puisque $\pi\left(s_B \mid s_A, b\right) = 1_\mathcal{L}$ et $\mu(s_B) = 1_\mathcal{L}$, il est clair que le revenu en $s_A$ pour un horizon 1 est $u_1^*(s_A) = 1_\mathcal{L}$. Évidemment, le revenu en $s_B$ pour un horizon 1 est $u_1^*(s_B) = 1_\mathcal{L}$, et puisque $\pi\left(s_A \mid s_A, \overline{a}\right) = 1_\mathcal{L}$, alors $\max_{s' \in \mathcal{S}} \min\left\{\pi\left(s' \mid s_A, a\right), u_1^*(s')\right\} = 1_\mathcal{L}$ quelle que soit l'action $a \in \left\{\overline{a}, b\right\} = \mathcal{A}$, c'est-à-dire toutes les actions sont optimales en $s_A$. La condition "if" de l'algorithme ci-dessouspermet de sélectionner l'action optimale $b$ pendant la première itération. Cette condition et l'initialisation, qui ne sont pas présents dans les algorithmes précédents de la littérature, sont nécessaires pour prouver l'optimalité de la stratégie.

La preuve, qui est assez longue et complexe, est présentée en Annexe A. Cet algorithme pour les $\pi$-PDM sera ensuite étendu aux $\pi$-PDMOM dans la prochaine section.

**Algorithme 1 :** Algorithme d'itération sur les revenus optimaux (IR)

**for** $s \in \mathcal{S}$ **do**
  $u^*(s) \leftarrow 0_{\mathcal{L}}$ ;
  $u^c(s) \leftarrow \mu(s)$ ;
  $\delta(s) \leftarrow \overline{a}$ ;
**while** $u^* \neq u^c$ **do**
  $u^* \leftarrow u^c$ ;
  **for** $s \in \mathcal{S}$ **do**
    $u^c(s) \leftarrow \max_{a \in \mathcal{A}} \max_{s' \in \mathcal{S}} \min \{ \pi(s' \mid s, a), u^*(s') \}$ ;
    **if** $u^c(s) > u^*(s)$ **then**
      $\delta(s) \in \operatorname{argmax}_{a \in \mathcal{A}} \max_{s' \in \mathcal{S}} \min \{ \pi(s' \mid s, a), u^*(s') \}$ ;
**return** $u^*$, $\delta$ ;

Comme mentionné par Sabbadin (1999), nous faisons l'hypothèse de l'existence d'une action "rester", notée $\overline{a}$, qui laisse le système dans le même état avec nécessité $1_{\mathcal{L}}$. Cette action peut être considérée comme l'équivalent possibiliste du facteur d'actualisation $\gamma$ dans le modèle probabiliste, puisqu'elle garantit la convergence de l'algorithme d'itération sur les revenus optimaux. Cependant, nous verrons que l'action $\overline{a}$ n'est utilisée au final que pour certains états satisfaisants. Notons que les processus déterministes peuvent faire l'objet d'une hypothèse similaire pour le calcul de stratégies dont la longueur n'est pas spécifiée (LaValle, 2006). L'ensemble de toutes les stratégies finies est noté $\Delta = \cup_{i \geqslant 1} \Delta_i$, et $\#\delta$ est la taille d'une stratégie $(\delta)$ en termes d'étapes de décision. Le critère optimiste pour un horizon infini peut maintenant être défini : si $(\delta) \in \Delta$,

$$u(s, (\delta)) = \max_{\tau \in \mathcal{T}_{\#\delta}} \min \{ \Pi(\tau \mid s, (\delta)), M(\tau) \}. \tag{5}$$

THÉORÈME 3. — *S'il existe une action $\overline{a}$ telle que, pour tout $s \in \mathcal{S}$, $\pi(s' \mid s, \overline{a}) = 1_{\mathcal{L}}$ si $s' = s$ et $0_{\mathcal{L}}$ sinon, alors l'algorithme 1 calcule le critère optimiste maximal et une stratégie optimale qui est stationnaire c'est-à-dire qui ne dépend pas de l'étape $t$ du processus.*

PREUVE. — Voir Annexe A. ■

Soit $s$ un état tel que $\delta(s) = \overline{a}$, où $\delta$ est la stratégie renvoyée par l'algorithme. Dans l'algorithme 1, $u^*(s)$ sera égal à $\mu(s)$ dès le premier passage dans la boucle while, et le restera toujours au cours des itérations : pour chaque $s' \in \mathcal{S}$, soit $\forall a \in \mathcal{A}$ $\mu(s) \geqslant \pi(s' \mid s, a)$, soit $\mu(s) \geqslant u^*(s')$. Si le problème est non trivial, celà signifie que $s$ est un but ($\mu(s) > 0_{\mathcal{L}}$) et que les degrés de possibilité de transition vers de meilleurs buts sont plus petits que le degré de préférence de $s$.

### *4.2. Algorithme d'itération sur les revenus optimaux pour les $\pi$-PDMOM*

L'algorithme d'Itération dans l'espace des Revenus pour les $\pi$-PDMOM peut maintenant être présenté. Afin de clarifier cet algorithme, posons

$$U(a, s'_v, o'_h, \beta_h) = \min \left\{ \beta^a(s'_v, o'_h), u^*(s'_v, \beta_h^{a,s'_v,o'_h}) \right\}.$$

Notons que l'algorithme 2 a la même structure que l'algorithme 1. Notons aussi qu'un $\pi$-PDMOM est un $\pi$-PDM sur l'espace d'état $\mathcal{S}_v \times \mathcal{B}_h^\pi$. Soit $s_v \in \mathcal{S}_v$, $\beta_h \in \mathcal{B}_h^\pi$ et maintenant $\Gamma_{\beta,\overline{a},s'_v}(\beta'_h) = \left\{ o'_h \in \mathcal{O}_h \mid \beta_h^{\overline{a},s'_v,o'_h} = \beta'_h \right\}$. Pour assurer l'hypothèse du théorème 3, il suffit d'assurer que $\max_{o'_h \in \Gamma_{\beta,\overline{a},s'_v}(\beta'_h)} \beta^{\overline{a}}(s'_v, o'_h) = 1_{\mathcal{L}}$ si $s'_v = s_v$ et $\beta'_h = \beta_h$, et $0_{\mathcal{L}}$ sinon. Cette propriété est vérifiée lorsque $\pi(s' \mid s, \overline{a}) = 1_{\mathcal{L}}$ si $s' = s$ (et $0_{\mathcal{L}}$ sinon) et il existe une observation "rien" $\overline{o}$ qui est reçue en tout état lorsque $\overline{a}$ est appliquée : $\pi(o' \mid s', \overline{a}) = 1_{\mathcal{L}}$ si $o' = \overline{o}$, et $0_{\mathcal{L}}$ sinon.

## 5. Résultats expérimentaux

Considérons un robot sur une grille de taille $g \times g$, avec $g > 1$. Il connait parfaitement sa position sur la grille $(x, y) \in \{1, \ldots, g\}^2$ à chaque étape du processus, ce qui constitue l'espace des états visibles $\mathcal{S}_v$. Il se trouve initialement à la position $s_{v,0} = (1, 1)$. Deux cibles immobiles sont présentes sur la grille : la "cible 1" est en $(x_1, y_1) = (1, g)$, la "cible 2" se trouve en $(x_2, y_2) = (g, 1)$ sur la grille, et le robot connait parfaitement leurs positions. Une des deux cibles est $A$, l'autre est $B$, et la mission du robot est d'identifier et d'atteindre $A$ aussitôt que possible. Le robot ne sait pas quelle cible est $A$ : les deux situations $A_1$ et $A_2$ correspondent respectivement à "la cible 1 est $A$" et "la cible 2 est $A$" et constituent l'espace des états cachés $\mathcal{S}_h$. Les actions $\mathcal{A}$ sont les déplacements dans les quatre directions ainsi que l'action "rester" ;

**Algorithme 2 :** Algorithme d'IR pour $\pi$-PDMOM

**for** $s_v \in \mathcal{S}_v$ **and** $\beta_h \in \mathcal{B}_h^\pi$ **do**
  $u^*(s_v, \beta_h) \leftarrow 0_{\mathcal{L}}$ ;
  $u^c(s_v, \beta_h) \leftarrow \mu(s_v, \beta_h)$ ;
  $\delta(s_v, \beta_h) \leftarrow \overline{a}$ ;
**while** $u^* \neq u^c$ **do**
  $u^* \leftarrow u^c$ ;
  **for** $s_v \in \mathcal{S}_v$ **and** $\beta_h \in \mathcal{B}_h^\pi$ **do**
    $u^c(s) \leftarrow \max_{a \in \mathcal{A}} \max_{s'_v \in \mathcal{S}} \max_{o'_h \in \mathcal{O}_h} U(a, s'_v, o'_h, \beta_h)$ ;
    **if** $u^c(s_v, \beta_h) > u^*(s_v, \beta_h)$ **then**
      $\delta(s) \in \operatorname{argmax}_{a \in \mathcal{A}} \max_{s'_v \in \mathcal{S}} \max_{o'_h \in \mathcal{O}_h} U(a, s'_v, o'_h, \beta_h)$ ;
**return** $u^*$, $\delta$ ;

les déplacements du robot sont déterministes. A chaque étape du processus, le robot analyse une image de chaque cible et obtient alors une observation de la nature de la cible : les deux cibles peuvent sembler être $A$ ($oAA$), ou bien seulement la cible 1 ($oAB$), ou seulement la cible 2 ($oBA$), ou alors aucune des deux ($oBB$).

Dans le cadre probabiliste, la probabilité de recevoir une bonne observation de la cible $i \in \{1, 2\}$, n'est pas vraiment connue, mais est approchée par

$$Pr\,(\,good_i \mid\ x, y\,) = \frac{1}{2}\left[1 + \exp\left(-\frac{\sqrt{(x-x_i)^2 + (y-y_i)^2}}{D}\right)\right]$$

où $(x, y) = s_v \in \{1, \ldots, g\}^2$ est la position du robot, $(x_i, y_i)$ la position de la cible $i$, et $D$ une constante de normalisation. Les processus d'observation de chaque cible sont considérés indépendants. Alors, par exemple, $Pr\,(\,oAB \mid (x, y), A_1\,)$ est égal à $Pr\,(\,good_1 \mid (x, y)\,) \cdot Pr\,(\,good_2 \mid (x, y)\,)$, $Pr\,(\,oAA \mid (x, y), A_1\,)$ à $Pr\,(\,good_1 \mid (x, y)\,) \cdot [1 - Pr\,(\,good_2 \mid (x, y)\,)]$, etc. Chaque étape du processus avant d'atteindre une cible coûte 1, atteindre la cible $A$ et y rester est récompensé par 100, et par $-100$ pour $B$. La stratégie provenant du modèle probabiliste a été calculée en tenant compte de l'observabilité mixte du problème, avec APPL (Ong *et al.*, 2010), en utilisant une précision de $0,046$ (la limite en mémoire est atteinte pour une précision supérieure) et $\gamma = 0,99$. Ce problème ne peut pas être résolu par l'algorithme exact pour PDMOM (Araya-Lòpez *et al.*, 2010) puisque cela entraîne l'utilisation de toute la mémoire vive disponible après 15 itérations.

Avec la théorie des possibilités qualitatives, on suppose qu'il est toujours possible d'observer correctement la cible : $\pi\,(\,good \mid\ x, y\,) = 1$. Ici, $\mathcal{L}$ sera un sous ensemble fini de $[0, 1]$, c'est pourquoi $1_{\mathcal{L}}$ peut être noté 1. Ensuite, plus le robot est loin de la cible $i$, plus il est susceptible de mal l'observer (par exemple observer $A$ au lieu de $B$), ce qui est une hypothèse raisonnable compte tenu du fait que le modèle d'observation est mal connu : $\pi\,(\,bad_i \mid x, y\,) = \frac{\sqrt{(x-x_i)^2+(y-y_i)^2}}{\sqrt{2}(g-1)}$. Ainsi, par exemple, $\pi\,(\,oAB \mid (x, y), A_1)$ est égal à 1, $\pi\,(\,oAA \mid (x, y), A_1)$ à $\pi\,(\,bad_2 \mid x, y\,)$, et $\pi\,(\,oBA \mid\ (x, y), A_1\,)$ à $\min\{\pi\,(\,bad_1 \mid\ x, y\,), \pi\,(\,bad_2 \mid\ x, y\,)\}$, etc. Puisque la situation est complètement connue lorsque le robot est sur la position d'une cible (observation déterministe), il n'y a pas de risque d'être bloqué dans un état non satisfaisant, et c'est pourquoi le modèle $\pi$-PDMOM *optimiste* fonctionne bien. L'échelle $\mathcal{L}$ est composée de 0, 1, et toutes les valeurs possibles de $\pi\,(\,bad_i \mid\ x, y\,) \in [0, 1]$. Notons qu'une construction de ce modèle avec une transformation probabilité-possibilité (Dubois *et al.*, 1993) aurait été équivalente. La distribution de préférence $\mu$ est égale à 0 pour tous les états du système, et à 1 pour les états $[(x_1, y_1), A_1]$ et $[(x_2, y_2), A_2]$ où $(x_i, y_i)$ est la position de la cible $i$. Comme mentionné par Sabbadin (1999), la stratégie calculée garantit un plus court chemin vers les états buts : la stratégie tend à réduire le temps de la mission.

Les algorithmes pour $\pi$-PDMPO standards, qui n'exploitent pas l'observabilité mixte contrairement à notre modèle $\pi$-PDMOM, ne peuvent pas résoudre le problème même pour de très petites grilles $3 \times 3$. En effet, pour ce problème, $\#\mathcal{L} = 5$, $\#\mathcal{S}_v = 9$,

et $\#\mathcal{S}_h = 2$. Ainsi, $\#\mathcal{S} = \#\mathcal{S}_v \cdot \#\mathcal{S}_h = 18$ et le nombre d'états de croyance est alors $\#\mathcal{B}^\pi = \mathcal{L}^{\#\mathcal{S}} - (\mathcal{L}-1)^{\#\mathcal{S}} = 5^{18} - 4^{18} \geqslant 3,7.10^{12}$ au lieu de 81 états avec un $\pi$-PDMOM. Par conséquent, les résultats expérimentaux qui suivent n'auraient pas pu être obtenus avec des $\pi$-PDMPO standards, ce qui justifie donc ce travail sur les $\pi$-PDMOM.

La comparaison des performances des modèles probabilistes et possibilistes est possible à l'aide des espérances de la somme des récompenses de leurs stratégies respectives : puisque la situation est complètement connue lorsque le robot est à la position d'une des cibles, il ne peut pas terminer en choisissant la cible $B$. Si $k$ est le nombre d'étapes du processus pour identifier et atteindre la bonne cible, alors la somme des récompenses est $100 - k$.

Considérons maintenant qu'en réalité (donc ici pour les simulations), et contrairement à ce qui est décrit par le modèle, la situation en pratique fait que l'algorithme de vision artificielle utilisé par le robot est trompeur lorsque le robot est loin des cibles, c'est-à-dire si quel que soit $i \in \{1,2\}$, $\sqrt{(x-x_i)^2+(y-y_i)^2} > C$, avec $C$ une constante positive, alors $Pr\,(\,good_i \mid\ x,y\,) = 1 - P_{bad} < \frac{1}{2}$. Dans tous les autres cas, le modèle probabiliste est effectivement le bon. La figure 3 résume le problème, et indique la zone où le robot a une mauvaise perception par la dénomination "error zone". Pour les expérimentations numériques qui suivent, le nombre de simulations était de $10^4$ pour calculer la moyenne de la somme des récompenses à l'exécution. La taille de la grille était de $10 \times 10$, $D = 10$ et $C = 4$.

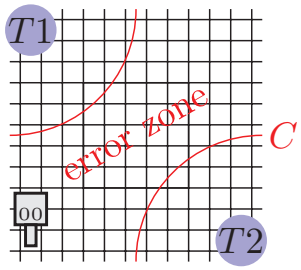



*Figure 3. Mission robotique de reconnaissance de cibles*

La figure 4.a montre que le modèle probabiliste est plus affecté par l'erreur introduite que le modèle possibiliste : elle représente la moyenne de la somme des récompenses à l'exécution obtenue par chaque modèle, comme une fonction de $P_{bad}$, la probabilité de mal observer une cible lorsque la position du robot est telle que $\sqrt{(x-x_i)^2+(y-x_i)^2} > C$. C'est dû au fait que la mise à jour possibiliste de l'état de croyance ne tient pas compte des nouvelles observations lorsque le robot en a déjà obtenu une plus fiable. Au contraire, le modèle probabiliste modifie l'état de croyance courant à chaque étape. En effet, puisqu'il n'y a que deux états cachés $s_h^1$ et $s_h^2$, si $\beta_h(s_h^1) < 1$, alors la normalisation possibiliste implique que $\beta_h(s_h^2) = 1$. La définition de la possibilité jointe de l'état et de l'observation (le minimum entre la distribution de possibilité sur l'état du système, c'est-à-dire l'état de croyance, et la possibilité de l'observation) assure que la possibilité jointe de $s_h^1$ et de l'observation

obtenue, est plus petite que $\beta_h(s_h^1)$. L'équivalent possibiliste de l'équation de mise à jour de l'état de croyance (3) assure donc que l'état de croyance suivant se retrouve dans un des trois cas suivant:

– elle est encore plus sceptique à propos de $s_h^1$ si l'observation est plus fiable, et confirme l'état de croyance précédent ($\pi\left(o_h \mid\ s_v, s_h^1, a\right)$ est plus petit que $\beta_h(s_h^1)$) ;

– elle devient l'état de croyance opposé si l'observation est plus fiable et contredit l'état de croyance précédent ($\pi\left(o_h \mid\ s_v, s_h^2, a\right)$ est à la fois plus petit que $\beta_h(s_h^1)$ et que $\pi\left(o_h \mid\ s_v, s_h^1, a\right)$) ;

– elle reste simplement la même si l'observation n'est pas plus informative que l'état de croyance courant.

Le théorème qui suit donne des conditions suffisantes menant à une mise à jour informative de l'état de croyance possibiliste. Classiquement un état de croyance $\beta_1 \in \mathcal{B}^\pi$ est dit plus spécifique qu'un état de croyance $\beta_2 \in \mathcal{B}^\pi$ si pour chaque $s \in \mathcal{S}$, $\beta_1(s) \leqslant \beta_2(s)$. Afin d'avoir un ordre total sur $\mathcal{B}^\pi$, l'échelle $\mathcal{L}$ est considérée ici comme un sous ensemble de $[0,1] \subset \mathbb{R}$. La relation d'ordre $\preceq$ sur $\mathcal{B}^\pi$ peut alors être définie pour classer les état de croyance selon leur spécificité :

$$\beta_1 \preceq \beta_2 \Leftrightarrow \sum_{s\in\mathcal{S}} \beta_1(s) \leqslant \sum_{s\in\mathcal{S}} \beta_2(s)$$

Notons que si $\beta_1$ est plus spécifique que $\beta_2$, alors $\beta_1 \preceq \beta_2$.

THÉORÈME 4. — *Soit $\beta_0 \in \mathcal{B}^\pi$ l'état de croyance initial modélisant l'ignorance totale, c'est-à-dire pour tous les $s \in \mathcal{S}$, $\beta_0(s) = 1$. Si la fonction de transition $T^\pi$ est déterministe, et si les observations ne sont pas informatives, c'est-à-dire $\forall s' \in \mathcal{S}$ , $\forall a \in \mathcal{A}$, $\forall o' \in \mathcal{O}$, $\Omega^\pi\left(s', a, o'\right) = 1$, alors si l'état de croyance $\beta_{t+1} \in \mathcal{B}^\pi$ est le résultat de la mise à jour de l'état de croyance $\beta_t \in \mathcal{B}^\pi$, $\beta_{t+1} \preceq \beta_t$. Ce résultat reste valable si pour chaque action $T^\pi$ est l'identité et $\forall (o', \tilde{o}) \in \mathcal{O}^2$, $\forall (a, \overline{a}) \in \mathcal{A}^2$ et $\forall s' \neq \tilde{s} \in \mathcal{S}$ t.q. $\Omega^\pi\left(s', a, o'\right) < 1_\mathcal{L}$, $\Omega^\pi\left(s', a, o'\right) \neq \Omega^\pi\left(\tilde{s}, \overline{a}, \tilde{o}\right)$.*

PREUVE. — Soit $\beta_t \in \mathcal{B}^\pi$. Rappelons tout d'abord la notation pour $a \in \mathcal{A}$ et $s' \in \mathcal{S}$, $\beta_t^a(s') = \max_{s\in\mathcal{S}} \min\left\{\pi\left(s' \mid\ s, a\right), \beta_t(s)\right\}$. Pour $s' \in \mathcal{S}$, notons $\mathcal{P}_a(s')$ l'ensemble des parents de $s'$: $\mathcal{P}_a(s') = \left\{s \in \mathcal{S} \mid\ \pi\left(s' \mid\ s, a\right) = 1_\mathcal{L}\right\}$. Comme $T^\pi$ est déterministe, $\left\{\mathcal{P}_a(s')\right\}_{s'\in\mathcal{S}}$ forme une partition de $\mathcal{S}$. En effet, $s \in \mathcal{S}$ a au moins un successeur, mais ne peut pas avoir plusieurs successeurs, c'est-à-dire $\forall s$, $\exists s' \in \mathcal{S}$ t.q. $s \in \mathcal{P}_a(s')$ et $\forall (s', \tilde{s}) \in \mathcal{S}^2$ $\mathcal{P}_a(s') \cap \mathcal{P}_a(\tilde{s}) = \emptyset$. Distinguons alors deux cas pour $s' \in \mathcal{S}$ :

– si $\mathcal{P}_a(s') = \emptyset$, alors $\beta_t^a(s) = 0_\mathcal{L}$ ;

– si $\mathcal{P}_a(s') \neq \emptyset$, alors $\beta_t^a(s) = \max_{s\in\mathcal{P}_a(s')} \beta_t(s)$.

Ainsi, puisque $\forall s' \in \mathcal{S}$, $\max_{s\in\mathcal{P}_a(s')} \beta_t(s) \leqslant \sum_{s\in\mathcal{P}_a(s')} \beta_t(s)$,

$$\sum_{s'\in\mathcal{S}} \beta_t^a(s') \leqslant \sum_{s'\in\mathcal{S}} \sum_{s\in\mathcal{P}_a(s')} \beta_t(s) = \sum_{s\in\mathcal{S}} \beta_t(s), \quad \text{c-à-d} \quad \beta_t^a \preceq \beta_t.$$

Finalement, puisque $\Omega^\pi(s', a, o') = 1_\mathcal{L}$, alors $\min\{\Omega^\pi(s', a, o'), \beta_t^a(s')\} = \beta_t^a(s')$, donc $\beta_t^a = \beta_{t+1}$, et $\beta_{t+1} \preceq \beta_t$, d'après l'équation de mise à jour de l'état de croyance (3).

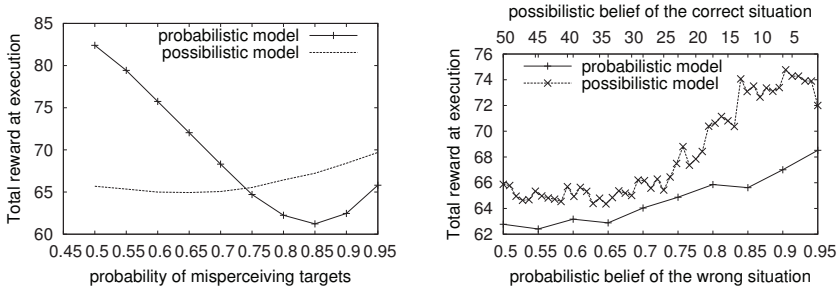
Nous montrons maintenant que ce résultat reste valable pour la seconde condition. Tout d'abord, comme $T^\pi$ est l'identité, $\beta_t^a = \beta_t$. Notons $\mathcal{L}_{s'} = \left\{\Omega^\pi(s', a, o') \middle| a \in \mathcal{A}, o' \in \mathcal{O}\right\}$. L'hypothèse sur la fonction $\Omega^\pi(s', a, o')$ permet de remarquer que $\mathcal{L}_{s'} \cap \mathcal{L}_{\tilde{s}} = \{1\}$, $\forall s' \neq \tilde{s}$. Ainsi, partant de l'état de croyance $\forall s \in \mathcal{S}$, $\beta_0(s) = 1_\mathcal{L}$, l'état de croyance suivant est $\forall s' \in \mathcal{S}$, $\beta_1(s') = \min\{\Omega^\pi(s', a_0, o_1), \beta_0(s')\} = \Omega^\pi(s', a_0, o_1) \in \mathcal{L}_{s'}$. Pour $t \geqslant 1$, la mise à jour (3) d'un état de croyance $\beta_t$ est la fonction $j_t : s' \mapsto \min\{\Omega^\pi(s', a_t, o_{t+1}), \beta_t(s')\}$ dont la plus grande valeur est remplacée par $1_\mathcal{L}$. Ainsi, si $\beta_t(s') \in \mathcal{L}_{s'}$, $\beta_{t+1}(s') \in \mathcal{L}_{s'}$: il est donc montré par récurrence que $\forall s' \in \mathcal{S}$, $t \geqslant 0$, $\beta_{t+1}(s') \in \mathcal{L}_{s'}$. De part la définition de $j_t$, $\forall s' \in \mathcal{S}$, $j_t(s') \in \mathcal{L}_{s'}$. Soit $\beta_t$ l'état de croyance à l'instant $t \geqslant 0$. Distinguons deux cas :

– si il existe $s^* \in \mathcal{S}$ t.q. $j_t(s') = 1_\mathcal{L}$, alors $\forall s' \in \mathcal{S}$, $\beta_{t+1}(s') = j_t(s')$. Ainsi, comme $j_t(s') = \min\{\Omega^\pi(s', a_t, o_{t+1}), \beta_t(s')\}$, $\beta_{t+1}$ est plus spécifique que $\beta_t$, et donc $\beta_{t+1} \preceq \beta_t$.

– sinon, $\forall s' \in \mathcal{S}$, $j_t(s') < 1_\mathcal{L}$. Cela implique qu'il existe un unique $s^*$ maximisant $j_t$. En effet, si cette fonction $j_t$ a la même valeur en deux états $s' \neq \tilde{s}$, alors $\mathcal{L}_{s'} \ni j_t(s') = j_t(\tilde{s}) \in \mathcal{L}_{\tilde{s}}$, et donc $\#\left(\mathcal{L}_{s'} \cap \mathcal{L}_{\tilde{s}}\right) > 1$, ce qui contredit la remarque initiale sur les ensembles $(\mathcal{L}_s)_{s\in\mathcal{S}}$. La mise à jour (3) mène alors à $\beta_{t+1}(s^*) = 1_\mathcal{L}$, et $\beta_{t+1}(s') = j_t(s')$ pour $s' \neq s^*$. Soit $s_* \in \mathcal{S}$ t.q. $\beta_t(s_*) = 1$. Nous avons $\beta_{t+1}(s_*) < \beta_t(s^*)$, car dans le cas contraire on aurait $\beta_{t+1}(s_*) \geqslant \beta_t(s^*) \geqslant \min\{\Omega^\pi(s^*, a_t, o_{t+1}), \beta_t(s^*)\} = j_t(s^*)$, et puisque $\beta_{t+1}(s_*) = j_t(s_*)$, $s^*$ ne maximiserait pas $j_t$ ce qui est une contradiction. Ainsi, $\sum_{s\in\mathcal{S}\setminus\{s^*,s_*\}} \beta_{t+1}(s) \leqslant \sum_{s\in\mathcal{S}\setminus\{s^*,s_*\}} \beta_t(s)$, et comme $\beta_{t+1}(s^*) = \beta_t(s_*) = 1_\mathcal{L}$ et $\beta_{t+1}(s_*) < \beta_t(s^*)$, nous avons bien $\beta_{t+1} \preceq \beta_t$.

■

La mise à jour probabiliste quant à elle ne permet pas à l'état de croyance de devenir directement l'état de croyance opposé, ou d'ignorer les observations moins fiables : le robot se dirige d'abord vers la mauvaise cible car il est initialement trop loin des deux cibles et les observe mal. Lorsqu'il est proche de cette cible, il reçoit de bonnes observations, et change petit à petit d'état de croyance : ce dernier devient assez informatif pour le convaincre de se diriger vers la cible $A$. Cependant, il passe alors inévitablement par la zone d'erreur : cela modifie peu à peu son état de croyance, qui devient faux avec grande probabilité, et le robot se retrouve dans la situation initiale. Il perd donc beaucoup de temps à sortir de cette boucle. On peut voir que la moyenne de la somme des récompenses croît lorsque la probabilité de mal observer $P_{bad}$ est très grande : cela s'explique par le fait que cette grande erreur mène le robot à atteindre la mauvaise cible plus rapidement, et donc à être quasiment sûr que la cible $A$ est l'autre cible.




(a) $P_{bad}$ varie (b) $\beta_0$ varie, $P_{bad} = 0,8$

*Figure 4. Comparaison des moyennes de la somme des récompenses à l'exécution, pour les modèles probabilistes et possibilistes.*

Maintenant, fixons $P_{bad} = 0,8$ et évaluons la moyenne de la somme des récompenses à l'exécution pour différents faux états de croyance initiaux : la figure 4.b illustre cette évaluation, avec les mêmes paramètres que pour la précédente expérimentation : nous comparons ici le modèle possibiliste, et la probabiliste lorsque l'état de croyance initial est fortement orienté vers la mauvaise cible (c'est-à-dire l'agent pense fortement que la cible 1 est $B$ alors que c'est $A$ en réalité). Notons que l'état de croyance possibiliste en la bonne cible décroît lorsque la nécessité en la mauvaise croît. Cette figure montre que le modèle possibiliste mène à de meilleures récompenses à l'exécution si l'état de croyance initial est mauvais et la fonction d'observation est imprécise: notons cependant que pour $\mathcal{P}_{bad} \leqslant 0,6$, la politique probabiliste est plus efficace [1].

## 6. Conclusion et perspectives

Nous avons proposé un algorithme d'Itération dans l'espace des Revenus (IR) pour les PDM possibilistes. Celui-ci calcule une stratégie optimale stationnaire pour un horizon infini contrairement aux méthodes précédentes. Une preuve complète de la convergence a été fournie : elle repose sur l'existence d'une action "rester" intermédiaire. Celle-ci est utilisée uniquement pour maintenir le processus dans les états buts. Enfin, cet algorithme a été étendu au nouveau modèle des PDM possibilistes qualitatifs à observabilité mixte, dont la complexité est exponentiellement plus petite que celle des PDMPO possibilistes qualitatifs. De ce fait, nous avons pu comparer les $\pi$-PDMOM avec leurs équivalents probabilistes sur un problème robotique réaliste. Nos résultats expérimentaux montrent que ces stratégies possibilistes peuvent être plus performantes que les stratégies issues du modèle probabiliste lorsque la fonction d'observation n'est pas connue précisément.

1. L'implémentation de l'algorithme de résolution, ainsi que la description de ce problème de reconnaissance qui en est l'entrée, sont disponibles sur le dépôt accessible à l'adresse https://github.com/drougui/ppudd : le problème peut être simulé en utilisant la stratégie optimale possibiliste calculée par l'algorithme.

Le cadre possibiliste qualitatif peut cependant être inadapté aux situations dans lesquelles quelques informations probabilistes sont disponibles à propos du problème : les PDMPO à paramètres imprécis (Itoh, Nakamura, 2007) et les PDMPO à paramètres bornés (Ni, Liu, 2012) intègrent la méconnaissance en considérant des ensembles de distributions de probabilité, plutôt qu'une seule. Lorsque de tels ensembles ne peuvent pas être déterminés, ou bien lorsqu'un modèle qualitatif suffit, les $\pi$-PDMPO peuvent être une bonne alternative : les PDMPOPI et les PDMPOPB sont très difficiles à résoudre en pratique.

La version pessimiste des $\pi$-PDM peut aussi être construite, mais l'optimalité de la stratégie renvoyée par l'algorithme d'IR associé semble difficile à prouver, essentiellement car il n'est pas suffisant de construire une trajectoire maximisante, comme fait dans la preuve en Annexe A. Les travaux de Weng (2007) et de Pralet *et al*. (2009) peuvent être utiles pour obtenir des résultats à propos des $\pi$-PDM pessimistes, afin de résoudre des problèmes contenant des situations dangereuses.

Enfin, ce formalisme a été étendu aux modèles factorisés dans (Drougard *et al*., 2014) : la version symbolique de l'algorithme 2 y est présentée, inspirée du travail de Hoey *et al*. (1999) sur les PDM probabilistes factorisés. Des arbres de décision algébriques sont utilisés pour représenter les paramètres des $\pi$-PDMOM au cours des itérations, et alléger les calculs.

*Remerciements*

*Nous tenons à remercier Régis Sabbadin, Hélène Fargier et Paul Weng pour nos échanges et leurs éclairages à propos des processus qualitatifs et des $\pi$-PDM.*

## Bibliographie

Araya-Lòpez M., Thomas V., Buffet O., Charpillet F. (2010). A closer look at MOMDPs. In *Proceedings of the twenty-second IEEE international conference on tools with artificial intelligence (ICTAI-10).*

Bellman R. (1957). A Markovian Decision Process. *Indiana Univ. Math. J.*, vol. 6, p. 679–684.

De Finetti B. (1937). La prévision: ses lois logiques, ses sources subjectives. In *Annales de l'Institut Henri Poincaré*, vol. 7, p. 1–68.

Drougard N., Teichteil-Königsbuch F., Farges J., Dubois D. (2014). Structured possibilistic planning using decision diagrams. In *Proceedings of the twenty-eighth AAAI conference on artificial intelligence, july 27 -31, 2014, Québec city, Québec, Canada.*, p. 2257–2263.

Dubois D., Fortemps P. (2005). Selecting preferred solutions in the minimax approach to dynamic programming problems under flexible constraints. *European Journal of Operational Research*, vol. 160, n° 3, p. 582-598.

Dubois D., Prade H. (1995). Possibility Theory as a basis for qualitative decision theory. In *Proceedings of the fourteenth international joint conference on artificial intelligence, IJCAI 95, Montréal Québec, Canada, August 20-25 1995, 2 volumes*, p. 1924–1932.

Dubois D., Prade H., Sandri S. (1993). On possibility/probability transformations. In R. Lowen, M. Roubens (Eds.), *Fuzzy logic: State of the art*, p. 103–112. Kluwer Academic Publ.

Dubois D., Prade H., Smets P. (1996). Representing partial ignorance. *IEEE Trans. on Systems, Man and Cybernetics*, vol. 26, p. 361–377.

Hoey J., St-aubin R., Hu A., Boutilier C. (1999). SPUDD: Stochastic planning using decision diagrams. In *Proceedings of the fifteenth conference on uncertainty in artificial intelligence*, p. 279–288. Morgan Kaufmann.

Itoh H., Nakamura K. (2007). Partially observable Markov decision processes with imprecise parameters. *Artificial Intelligence*, vol. 171, n° 8–9, p. 453 - 490.

LaValle S. M. (2006). *Planning Algorithms*. New York, NY, USA, Cambridge University Press.

Ni Y., Liu Z.-Q. (2012). Policy iteration for bounded-parameter POMDPs. *Soft Computing*, vol. 17, p. 1-12.

Ong S. C. W., Png S. W., Hsu D., Lee W. S. (2010, juillet). Planning under uncertainty for robotic tasks with mixed observability. *Int. J. Rob. Res.*, vol. 29, n° 8, p. 1053–1068.

Pralet C., Schiex T., Verfaillie G. (2009). *Sequential decision-making problems - representation and solution*. Wiley.

Puterman M. L. (1994). *Markov decision processes: Discrete Stochastic Dynamic Programming* (1st éd.). New York, NY, USA, John Wiley & Sons, Inc.

Sabbadin R. (1999, juillet). A possibilistic model for qualitative sequential decision problems under uncertainty in partially observable environments. In *15th Conference on Uncertainty in Artificial Intelligence (UAI99) , Stockholm, 30/07/99-01/08/99*, p. 567–564. San Francisco, Morgan Kaufmann.

Sabbadin R. (2000). Empirical comparison of probabilistic and possibilistic markov decision processes algorithms. In *ECAI 2000, proceedings of the 14th european conference on artificial intelligence, berlin, germany, august 20-25, 2000*, p. 586–590.

Sabbadin R. (2001). Possibilistic Markov decision processes. *Engineering Applications of Artificial Intelligence*, vol. 14, n° 3, p. 287 - 300. (Soft Computing for Planning and Scheduling)

Sabbadin R., Fargier H., Lang J. (1998). Towards qualitative approaches to multi-stage decision making. *International Journal of Approximate Reasoning*, vol. 19, n° 3–4, p. 441 - 471.

Smallwood R., Sondik E. (1973). *The Optimal Control of Partially Observable Markov Processes Over a Finite Horizon* (vol. 21). INFORMS.

Weng P. (2007). Conditions générales pour l'admissibilité de la programmation dynamique dans la décision séquentielle possibiliste. *Revue d'Intelligence Artificielle*, vol. 21, n° 1, p. 129-143. (NAT LIP6 DECISION)

## Annexe A. Démonstration du théorème 3

Cette annexe prouve que l'algorithme 1 renvoie le revenu maximal du critère de l'équation (5) et une stratégie optimale. Nous rappelons qu'il existe une action notée $\overline{a} \in \mathcal{A}$ telle que pour chaque $s \in \mathcal{S}$, $\pi(s' \mid s, \overline{a}) = 1_{\mathcal{L}}$ si $s' = s$, et $0_{\mathcal{L}}$ sinon. L'existence de cette action $\overline{a}$ implique que le revenu est croissant par rapport à la taille de l'horizon, c'est-à-dire le critère (1) croît avec $p$ :

LEMME 5. — $\forall s \in \mathcal{S}, \forall p \geqslant 0, u^*_p(s) \leqslant u^*_{p+1}(s)$.

PREUVE. — $u^*_{p+1}(s_0) = \max_{\Delta_{p+1}} \max_{\tau \in \mathcal{T}_{p+1}} \min \left\{ \min_{i=0}^{p} \pi(s_{i+1} \mid s_i, \delta_i(s_i)), \mu(s_{p+1}) \right\}$, pour un état $s_0 \in \mathcal{S}$. Considérons les trajectoires $\tau' \in \mathcal{T}'_{p+1} \subset \mathcal{T}_{p+1}$ telles que $\tau' = (s_1, \ldots, s_p, s_p)$. La règle de décision $\overline{\delta}$ est telle que pour chaque $s \in \mathcal{S}$, $\overline{\delta}(s) = \overline{a}$. Considérons aussi les stratégies $(\delta') \in \Delta'_{p+1} \subset \Delta_{p+1}$ telles que $(\delta') = (\delta_0, \ldots, \delta_{p-1}, \overline{\delta})$. Il est évident que

$$u^*_{p+1}(s_0) \geqslant \max_{(\delta') \in \Delta'_{p+1}} \max_{\tau' \in \mathcal{T}'_{p+1}} \min \left\{ \min_{i=0}^{p} \pi(s_{i+1} \mid s_i, \delta_i(s_i)), \mu(s_{p+1}) \right\}.$$

Notons que le membre droit de cette inégalité peut se réécrire

$$\max_{(\delta) \in \Delta_p} \max_{\tau \in \mathcal{T}_p} \min \left\{ \min_{i=0}^{p-1} \pi(s_{i+1} \mid s_i, \delta_i(s_i)), \pi(s_p \mid s_p, \overline{a}), \mu(s_p) \right\}$$

$= u^*_p(s_0)$ puisque $\pi(s_p \mid s_p, \overline{a}) = 1_{\mathcal{L}}$. ■

La signification de ce lemme est qu'il est toujours plus possible d'atteindre un état $s$ à partir de $s_0$ en au plus $p+1$ étapes qu'en au plus $p$ étapes. Puisque pour chaque $s \in \mathcal{S}$, $(u^*_p(s))_{p \in \mathbb{N}} \leqslant 1_{\mathcal{L}}$, le lemme 5 assure que la suite $(u^*_p(s))_{p \in \mathbb{N}}$ converge. L'étape suivante montre que la convergence de cette suite se produit en temps fini.

LEMME 6. — *Pour tout $s \in \mathcal{S}$, le nombre d'itérations de la suite $(u^*_p(s))_{p \in \mathbb{N}}$ avant la convergence est borné par $\#\mathcal{S} \cdot \#\mathcal{L}$.*

PREUVE. — Rappelons tout d'abord que les degrés de possibilité et de préférence sont dans $\mathcal{L}$ qui est fini et totalement ordonné : nous pouvons écrire $\mathcal{L}$=$\{0_{\mathcal{L}}, l_1, l_2, \ldots, 1_{\mathcal{L}}\}$ avec $0_{\mathcal{L}} < l_1 < l_2 < \ldots < 1_{\mathcal{L}}$. Si les fonctions successives $u^*_k$ et $u^*_{k+1}$ sont égales, alors $\forall s \in \mathcal{S}$, les suites $(u^*(s))_{p \geqslant k}$ sont constantes. En effet, cette suite peut être définie par la formule de récurrence

$$u^*_p(s) = \max_{a \in \mathcal{A}} \max_{s' \in \mathcal{S}} \min \left\{ \pi(s' \mid s, a), u^*_{p-1}(s') \right\}.$$

Donc si $\forall s \in \mathcal{S}$, $u^*_p(s) = u^*_{p-1}(s)$ alors l'itération suivante $(p+1)$ fait face à la même situation ($u^*_{p+1}(s) = u^*_p(s)$ $\forall s \in \mathcal{S}$). La convergence la plus lente peut alors être décrite comme suit : pour chaque $p \in \mathbb{N}$ seulement un $s \in \mathcal{S}$ est tel que $u^*_{p+1}(s) > u^*_p(s)$. De plus, pour cet $s$, si $u^*_p(s) = l_i$, alors $u^*_{p+1}(s) = l_{i+1}$. Nous pouvons conclure que pour $p > \#\mathcal{L} \cdot \#\mathcal{S}$, la suite est constante. ■

Notons que la variable $u^*(s)$ de l'algorithme 1 est égale à $u_p^*(s)$ après l'itération $p$. Nous pouvons conclure que $u^*$ converge vers le revenu maximal du critère pour un horizon de taille $(\#\mathcal{L} \cdot \#\mathcal{S})$, et ne peut pas être plus grand : la fonction $u^*$ renvoyée est donc optimale par rapport à l'équation (5) et est calculée en un nombre fini d'itérations.

Dans la suite, nous prouvons l'optimalité de la stratégie $(\delta^*)$ renvoyée par l'algorithme 1. Pour cela, nous allons construire une trajectoire de taille plus petite que $\#\mathcal{S}$ qui maximise $\min\{\Pi(\tau \mid s_0, (\delta)), M(\tau)\}$ avec la stratégie $(\delta^*)$. Les deux prochains lemmes sont nécessaires pour cette construction, ainsi que quelques notations.

Soit $s_0 \in \mathcal{S}$ et $p$ le plus petit entier tel que $\forall p' \geqslant p$, $u_{p'}^*(s_0) = u^*(s_0)$, où $u^*$ est ici la valeur maximale du critère en horizon infini de l'équation (5) : la variable $u^*(s)$ de l'algorithme 1 n'augmente plus après $p$ itérations. L'équation (2) peut être utilisée pour renvoyer une stratégie optimale (non stationnaire) notée $(\delta^{(s_0)}) \in \Delta_p$. Avec cette notation : $\forall s \in \mathcal{S}, \quad \delta^*(s) = \delta_0^{(s)}(s)$. Considérons maintenant une trajectoire $\tau = (s_1, s_2, \ldots, s_p)$ qui maximise $\min\left\{\min_{i=0}^{p-1} \pi\left(s_{i+1} \mid s_i, \delta_i^{(s_0)}(s_i)\right), \mu(s_p)\right\}$. Cette trajectoire est appelée *trajectoire optimale de taille minimale partant de* $s_0$.

LEMME 7. — *Soit $\tau = (s_1, \ldots, s_p)$ une trajectoire optimale de taille minimale partant de $s_0$. Alors, $\forall k \in \{1, \ldots, p-1\}$, $u^*(s_0) \leqslant u^*(s_k)$.*

PREUVE. — Soit $k \in \{1, \ldots, p-1\}$.

$$\begin{aligned} u^*(s_0) &= \min\left\{\min_{i=0}^{p-1} \pi\left(s_{i+1} \mid s_i, \delta_i^{(s_0)}(s_i)\right), \mu(s_p)\right\} \\ &\leqslant \min\left\{\min_{i=k}^{p-1} \pi\left(s_{i+1} \mid s_i, \delta_i^{(s_0)}(s_i)\right), \mu(s_p)\right\} \leqslant u_{p-k}^*(s_k) \leqslant u^*(s_k) \end{aligned}$$

puisque $(u_p^*)_{p\in\mathbb{N}}$ est croissante d'après le lemme 5. ■

LEMME 8. — *Soit $\tau = (s_1, \ldots, s_p)$ une trajectoire optimale de taille minimale partant de $s_0$ et $k \in \{1, \ldots, p-1\}$. Si $u^*(s_0) = u^*(s_k)$, alors $\delta^*(s_k) = \delta_k^{(s_0)}(s_k)$.*

PREUVE. — Supposons que $u^*(s_0) = u^*(s_k)$. Puisque $u^*(s_0) \leqslant u_{p-k}^*(s_k) \leqslant u^*(s_k)$ d'après le lemme 7, nous obtenons que $u_{p-k}^*(s_k) = u^*(s_k)$. Le critère en $s_k$ est donc optimisé en un horizon $(p-k)$.

De plus un horizon plus petit ne mène pas à l'optimalité : $\forall m \in \{1, \ldots, p-k\}$, $u_{p-k-m}^*(s_k) < u^*(s_k)$ c'est-à-dire avec un horizon $p-k-m$ le critère en $s_k$ n'est pas maximisé. Dans le cas contraire, le critère en $s_0$ serait maximisé avec un horizon $(p-m)$ : la stratégie

$$\delta' = (\delta_0^{(s_0)}, \delta_1^{(s_0)}, \ldots, \delta_{k-1}^{(s_0)}, \delta_0^{(s_k)}, \ldots, \delta_{p-k-m-1}^{(s_k)}) \in \Delta_{p-m}$$

aurait été optimale. En effet,

$$\begin{aligned} u^*(s_0) &= \min\left\{ \min_{i=0}^{k-1} \pi\left( s_{i+1} \mid s_i, \delta_i^{(s_0)}(s_i)\right), u^*_{p-k}(s_k)\right\} \\ &= \min\left\{ \min_{i=0}^{k-1} \pi\left( s_{i+1} \mid s_i, \delta_i^{(s_0)}(s_i)\right), u^*(s_k)\right\} \end{aligned}$$

Soit $\overline{\tau} = (\overline{s}_1, \ldots, \overline{s}_{p-k-m}) \in \mathcal{T}_{p-k-m}$ une trajectoire optimale de taille minimale partant de $s_k$. En posant $\overline{s_0} = s_k$, $\overline{\tau}$ maximise donc

$$u^*(s_k) = \min\left\{ \min_{i=0}^{p-k-m-1} \pi\left( \overline{s}_{i+1} \mid \overline{s}_i, \delta_i^{(s_k)}(\overline{s}_i)\right), \mu(\overline{s}_{p-k-m})\right\}.$$

Si $(s'_1, \ldots, s'_{p-m}) = (s_1, \ldots, s_{k-1}, \overline{s}_0, \ldots, \overline{s}_{p-k-m})$,

$$u^*(s_0) = \min\left\{ \min_{i=0}^{p-m-1} \pi\left( s'_{i+1} \mid s'_i, \delta'_i(s_i)\right), \mu(s'_{p-m})\right\}$$

c-à-d $\exists p' = p - m < p$ tel que $u^*(s_0) = u^*_{p'}(s_0)$ : cela contredit l'hypothèse selon laquelle $(s_1, \ldots, s_p)$ est une trajectoire optimale de taille minimale.

Donc $p - k$ est bien le plus petit entier tel que $u^*_{p-k}(s_k) = u^*(s_k)$ : nous pouvons finalement conclure que $\delta^*(s_k)$ $(:= \delta_0^{(s_k)}(s_k))$ $= \delta_k^{(s_0)}(s_k)$. En effet, le dernier passage dans la condition *if* de l'algorithme 1 pour l'état $s_k$ se produit à l'itération $p - k$. ■

THÉORÈME 9. — *Soit $(\delta^*)$ la stratégie renvoyée par l'algorithme 1 ; $\forall s_0 \in \mathcal{S}$, il existe $p^* \leqslant \#\mathcal{S}$ et une trajectoire $(s_1, \ldots, s_{p^*})$ telle que*

$$u^*(s_0) = \min\left\{ \min_{i=0}^{p^*-1} \pi\left( s_{i+1} \mid s_i, \delta^*(s_i)\right), \mu(s_{p^*})\right\} :$$

*c'est-à-dire $\delta^*$ est une stratégie optimale.*

PREUVE. — Soit $s_0$ un état dans $\mathcal{S}$ et $\tau$ une trajectoire optimale de taille minimale $p$ partant de $s_0$. Si $\forall k \in \{1, \ldots, p-1\}$, $u^*(s_k) = u^*(s_0)$, alors d'après le lemme 8, $\delta^*(s_k) := \delta_0^{(s_k)}(s_k) = \delta_k^{(s_0)}(s_k)$, et donc le critère en $s_0$ est maximisé par $(\delta^*)$ puisque il est maximisé par $(\delta^{(s_0)})$ : l'optimalité est montrée.

Sinon, soit $k$ le plus petit entier dans $\{1, \ldots, p-1\}$ tel que $u^*(s_k) > u^*(s_0)$. La définition de $k$ assure que $u^*(s_k) > u^*(s_i)$ $\forall i \in \{0, \ldots, k-1\}$.

En réitérant et partant de $s_0^{(1)} = s_k$, soit $p^{(1)}$ le nombre d'itérations avant que la variable $u^*(s^{(1)})$ de l'algorithme ne converge (le plus petit entier tel que $u^*(s_0^{(1)}) = u^*_{p^{(1)}}(s_0^{(1)})$); Soit $\tau^{(1)} \in T_{p^{(1)}}$ la trajectoire qui maximise $\min\{\min_{i=0}^{p^{(1)}-1} \pi(s_{i+1}|s_i, \delta_i^{(s_0^{(1)})}(s_i)), \mu(s_p^{(1)})\}$ : $\tau^{(1)}$ est une trajectoire optimale de taille minimale partant de $s_k = s_0^{(1)}$. Nous sélectionnons $k^{(1)}$ de la même manière que précédemment, et réitérons en commençant par $s_0^{(2)} = s_{k^{(1)}}^{(1)}$ qui est

tel que $u^*(s^{(1)}_{k^{(1)}}) > u^*(s^{(1)}_0)$, et $u^*(s^{(1)}_{k^{(1)}}) > u^*(s^{(1)}_i)\ \forall i \in \{0,\dots,k^{(1)}-1\}$, etc... Le lemme 10, plus bas, montre que les états sélectionnés comme décrit, $(s_0,\dots,s_{k-1},s^{(1)}_0,\dots,s^{(1)}_{k^{(1)}-1},s^{(2)}_0\dots,s^{(2)}_{k^{(2)}-1},s^{(3)}_0,\dots)$, sont tous différents. Donc ce processus de sélection termine puisque $\#\mathcal{S}$ est un ensemble fini. Le nombre total d'états sélectionnés est noté $p^* = k + \sum_{i=1}^{q-1} k^{(i)} + p^{(q)}$ avec $q \geqslant 0$ le nombre de nouvelles trajectoires sélectionnées. Ainsi, la stratégie $(\delta') = (\delta_0,\dots,\delta_{k-1},\delta^{(s^{(1)}_0)}_0,\dots,\delta^{(s^{(1)}_0)}_{k^{(1)}-1},\dots,\delta^{(s^{(q)}_0)}_{p^{(q)}})$ correspond à $(\delta^*)$ sur $\tau' = (s'_1,\dots,s'_{p^*}) = (s_0,s_1,\dots,s_{k-1},s^{(1)}_0,\dots,s^{(1)}_{k^{(1)}-1},\dots,s^{(m)}_{p^{(q)}-1})$ et cette stratégie est optimale puisque $u^*(s_0) = u(s_0,(\delta^*))$ :

$$
\begin{aligned}
u^*(s_0) &= \min\left\{ \min_{i=0}^{k-1} \pi\left(s'_{i+1} \,\middle|\, s'_i, \delta'(s'_i)\right), u^*_{p-k}(s_k) \right\} \\
&\leqslant \min\left\{ \min_{i=0}^{k-1} \pi\left(s'_{i+1} \,\middle|\, s'_i, \delta'(s'_i)\right), u^*(s_k) \right\} \\
&= \min\left\{ \min_{i=0}^{k^{(1)}-1} \pi\left(s'_{i+1} \,\middle|\, s'_i, \delta'(s'_i)\right), u^*_{p^{(1)}-k^{(1)}}(s_{k^{(1)}}) \right\} \\
&\dots \leqslant \min\left\{ \min_{i=0,\dots,p^*-1} \pi\left(s'_{i+1} \,\middle|\, s'_i, \delta'(s'_i)\right), \mu(s'_{p^*}) \right\}
\end{aligned}
$$

Les signes "$\leqslant$" sont en fait des "$=$" puisque sinon, nous aurions trouvé une stratégie telle que $u(s_0,(\delta')) > u^*(s_0)$. Donc $(\delta^*)$ est optimale : $u^*(s_0) = \min\left\{ \min_{i=0}^{p^*-1} \pi\left(s'_{i+1} \,\middle|\, s'_i, \delta^*(s'_i)\right), \mu(s'_{p^*}) \right\}$ ■

COROLLAIRE 10. — *Le processus décrit dans la preuve précédente, afin de construire une trajectoire maximisant le critère avec $(\delta^*)$, sélectionne toujours des états différents.*

PREUVE. — Tout d'abord, deux états égaux dans la même trajectoire sélectionnée $\tau^{(m)}$ serait une contradiction de l'hypothèse que $p^{(m)}$ est le plus petit entier tel que $u^*_{p^{(m)}}(s^{(m)}_0) = u^*(s^{(m)}_0)$. En effet, soit $k$ et $l$ tels que $0 \leqslant k < l \leqslant p^{(m)}$ et supposons que $s^{(m)}_k = s^{(m)}_l$. Pour plus de clarté dans les explications qui vont suivre, nous omettons le "$(m)$" : $p = p^{(m)}$ et $\forall i \in \{0,\dots,l\}$, $s_i = s^{(m)}_i$.

$$
u^*_{p-k}(s_k) = \min\left\{ \min_{i=k}^{l-1} \pi\left(s_{i+1} \,\middle|\, s_i, \delta^{(s_0)}_i(s_i)\right), u^*_{p-l}(s_l) \right\} \leqslant u^*_{p-l}(s_l) = u^*_{p-l}(s_k).
$$

Cependant $u^*_{p-l}(s_k) \leqslant u^*_{p-k}(s_k)$ car la suite est croissante et $p-l < p-k$. Nous obtenons finalement que $u^*_{p-k}(s_k) = u^*_{p-l}(s_k)$, et donc

$$\begin{aligned} u^*(s_0) &= \min\left\{ \min_{i=0}^{k-1} \pi\left( s_{i+1} \mid s_i, \delta_i^{(s_0)}(s_i) \right), u^*_{p-k}(s_k) \right\} \\ &= \min\left\{ \min_{i=0}^{k-1} \pi\left( s_{i+1} \mid s_i, \delta_i^{(s_0)}(s_i) \right), u^*_{p-l}(s_l) \right\} \\ &= \min\left\{ \min_{i=0,\ldots,k-1,l,\ldots,p-1} \pi\left( s_{i+1} \mid s_i, \delta_i^{(s_0)}(s_i) \right), \mu(s_p) \right\} \end{aligned}$$

En conséquence, un horizon $(p^{(m)} - l + k)$ est suffisant pour atteindre le revenu optimal : c'est une contradiction avec la définition de $p^{(m)}$ puisque $p^{(m)} - l + k < p^{(m)}$.

Finalement, si nous supposons qu'un état $\overline{s}$ apparaît deux fois dans la suite complète d'états sélectionnée, alors cet état appartient à deux trajectoires $\tau^{(m)}$ et $\tau^{(m')}$ (avec $m < m'$). Le lemme 7 et la définition de $k^{(m)}$ qui implique que $u^*(s_0^{(m+1)})$ est strictement plus grand que le revenu optimal du critère en chacun des états $s_0^{(m)}, \ldots, s_{k^{(m)}-1}^{(m)}$, nécessite que $u^*(s_0^{(m')}) \leqslant u^*(\overline{s}) < u^*(s_0^{(m+1)})$. C'est une contradiction car $u^*(s_0^{(m+1)}) \leqslant u^*(s_0^{(m')})$ puisque $m < m'$. ■