Optimisation de POMDP : quelles récompenses sont réellement attendues à l'exécution de la politique ?

Caroline P. Carvalho Chanel[1,2], Jean-Loup Farges[2],
Florent Teichteil-Königsbuch[2], Guillaume Infantes[2]

[1] ISAE - Institut Supérieur de l'Aéronautique et de l'Espace
10, av. Edouard Belin BP 54032 - 31055 Toulouse Cedex 4

[2] ONERA - Office National d'Études et de Recherches Aérospatiales
2, av. Edouard Belin BP 4025 - 31055 Toulouse Cedex 4
prenom.nom@onera.fr

Résumé : Les Processus Décisionnels Markoviens Partiellement Observables sont actuellement un sujet d'intérêt dans la communauté scientifique grâce aux progrès observés dans des algorithmes de résolution et dans les capacités numériques de calcul. La plupart de ces algorithmes sont focalisés sur la résolution d'un critère de performance, qui a pour ambition de caractériser les politiques qui permettront de générer les séquences de récompenses le plus importantes possibles. Dans la planification en Intelligence Artificielle, l'attention est tournée vers un critère qui optimise une somme pondérée des récompenses, et, pour des applications en perception active d'autre part, le critère est souvent défini en termes de gain d'information (entropie de Shannon). Aucun de ces critères ne prend en compte les récompenses réellement acquises lors de l'exécution de la politique. En effet, le premier critère est une moyenne linéaire sur l'espace d'états de croyance, de sorte que l'agent ne tend pas à obtenir une meilleure information des différentes observations, alors que le second critère ne prend pas en compte les récompenses. Ainsi, motivés par des exemples démonstratifs, nous étudions deux combinaisons, additive et multiplicative, de ces critères afin d'obtenir une meilleur séquence de récompenses et de gain d'information lors de l'exécution de la politique. Nous comparons nos critères avec le critère classique optimisé (γ-*pondéré*) dans le cadre POMDP et nous soulignons l'intérêt de considérer un nouveau critère hybride non-linéaire pour des applications réalistes de reconnaissance et pistage multi-cibles.

1 Introduction

La plupart des applications réalistes en intelligence artificielle (IA) requièrent de planifier des actions avec une information incomplète de l'état du monde. Par exemple, un robot a besoin de trouver son chemin jusqu'à un but sans connaissance parfaite du monde qui l'entoure, ni sa localisation parfaite dans ce monde. Comme autre exemple, un agent contrôlant une caméra doit planifier les meilleur prises d'images en jouant sur les paramètres optiques d'orientation physique de la caméra pour identifier précisément un objet le plus rapidement possible. Si les effets des actions et des observations sont probabilistes, les Processus Décisionnels Markoviens Partiellement Observables (POMDP) sont un modèle expressif, mais longtemps négligé (du fait de la complexité prohibitive) pour la décision séquentielle avec observation partielle de l'état de l'environnement (Kaelbling *et al.*, 1998). Néanmoins, de récents progrès dans des algorithmes de résolution des POMDP (Pineau *et al.*, 2003; Spaan & Vlassis, 2005; Sridharan *et al.*, 2008) ont relancé une recherche intensive sur les algorithmes et les applications des POMDP.

Un POMDP est défini comme un n-uplet $\langle S, A, \Omega, T, O, R, b_0 \rangle$ où : S est un ensemble d'états, A un ensemble d'actions, Ω un ensemble d'observations, $T : S \times A \times S \to [0; 1]$ une fonction de transition entre les états : $T(s_t, a, s_{t+1}) = P(s_{t+1} \mid a, s_t)$; $O : \Omega \times S \to [0; 1]$ une fonction d'observation : $O(o_t, s_t) = P(o_t|s_t)$; $R : S \times A \times S \to \mathbb{R}$ une fonction de récompenses associées aux transitions, et b_0 une distribution de probabilités sur les états initiaux. On note B l'ensemble des distributions de probabilités sur les états, appelé aussi espace d'états de croyance. À chaque pas de temps t, l'agent met à jour son *état de croyance* défini en tant qu'élément $b_t \in B$.

L'objectif de la résolution d'un POMDP est de construire une politique, c'est-à-dire une fonction $\pi : B \to A$ qui maximise un critère généralement basé sur les récompenses linéarisées sur les états de croyance.

Dans la robotique, où des récompenses symboliques doivent être attendues, il est généralement accepté d'optimiser à long terme l'espérance de la somme pondérée des récompenses pour tout état de croyance initial (Cassandra *et al.*, 1996; Spaan & Vlassis, 2004) :

$$V^{\pi}(b) = E_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(b_t, \pi(b_t)) \mid b_0 = b\right]$$

Suivant les théorèmes d'optimalité, la fonction de valeur optimale des états de croyance est linéaire pas morceaux et convexe (Sondik & LABS, 1971), ce qui offre un modèle mathématique relativement simple pour le raisonnement, sur lequel la plupart des algorithmes sont basés. Toutefois, comme souligné et exploré dans cet article, la linéarisation de la valeur moyenne sur les états de croyance revient à aplatir les résultats des observations et finalement à perdre des informations distinctives à leur sujet. Pour cette raison, la politique optimisée ne pousse pas l'agent à acquérir suffisamment d'information sur son environnement avant d'agir pour gagner des récompenses : comme souligné dans cet article, une telle stratégie résulte dans la collecte de moindres récompenses lors de l'exécution que ce qui était prévu, particulièrement si l'état de croyance de départ est erroné.

Ce point crucial qui prête à confusion mérite plus d'explications pour une meilleur compréhension du sujet principal de cet article. Premièrement, il peut paraître étrange que la politique qui maximise la somme de récompenses n'ait pas le comportement espéré lors de son exécution : quel est le rapport entre le critère optimisé et les récompenses réellement gagnées lors de l'exécution ? Pourquoi peuvent-ils être aussi différents ?

En effet, le critère γ*-pondéré* est défini sur les états de croyance (l'agent applique une politique basée uniquement sur son état de croyance), alors que les récompenses obtenues lors de l'exécution sont cumulées sur la base des états actuels et successifs du système, cachés de l'agent. Avec une observabilité totale (cas des MDP), ce type de problème ne se présente pas, car les états réels du système et de l'agent sont complètement connus, et donc le critère est moyenné sur les chemins probabilistes réels définis par la fonction de transition. Mais dans le cas partiellement observable (POMDP), le critère est *moyenné sur un chemin probabiliste défini sur l'état de croyance*, lequel se présente généralement différemment du chemin d'exécution réel.

Curieusement, cette relation entre le critère optimisé et les récompenses réelles obtenues n'est pas très étudié : à notre connaissance, la recherche robotique pour les POMDP essaie presque exclusivement de trouver un moyen de plus en plus efficace d'optimiser l'espérance de la somme pondérée des récompenses sans tenir compte de, ou sans s'interroger sur, l'absence de séparation claire entre les actions d'observations éventuelles lors de l'optimisation.

D'autre part, les recherches en perception active visent à maximiser la quantité d'information obtenue de l'environnement (Deinzer *et al.*, 2003; Eidenberger *et al.*, 2008; Paletta & Pinz, 2000), en minimisant souvent l'entropie de Shannon, qui peut renseigner sur la quantité d'information contenue dans l'état de croyance. Comme exemple de critère utilisé, pour un b_0 donné on a : $H(b_0) = \sum_{t=0}^{+\infty} \gamma^t \sum_{s \in S} b_t(s).\log(b_t(s))$. Contrairement au critère vu plus haut, celui-ci est non-linéaire sur les états de croyance, et permet une distinction claire entre les observations pour effectuer une mise à jour de l'état de croyance dans la bonne direction. Mais ce critère ne prend pas en compte les récompenses éventuellement associées à un but de mission à atteindre.

Partant des approches de ces deux communautés, il semble naturel de rechercher un nouveau critère d'optimisation non-linéaire basé sur l'agrégation dans un seul critère des récompenses et de l'entropie d'état de croyance. De cette façon, les politiques optimisées consistent à alterner de l'acquisition d'informations et des actions qui mèneront l'agent au but de mission tout en maximisant la collecte de récompenses lors de l'exécution, car l'entropie jouera un rôle pénalisant pour les grandes incertitudes (en supposant que les deux critères soient bien contrebalancés). Formellement, notant $J_\lambda(V, H)$ comme un critère mixte qui dépend d'un paramètre $\lambda \in \Lambda$, le problème général que nous posons ici peut être formalisé par :

$$\max_{\lambda \in \Lambda} E\left[\sum_{t=0}^{+\infty} \gamma^t r_t \mid s_0, \pi_\lambda\right] \text{ où } \pi_\lambda = \operatorname*{argmax}_{\pi \in A^S} J_\lambda(V(b_0), H(b_0))$$

Autrement dit, quelle est la valeur de λ qui permet d'équilibrer la relation entre $V(b_0)$ et $H(b_0)$ afin d'augmenter l'espérance de la somme des récompenses gagnées effectivement pendant l'*exécution*, quand un critère mixte basé sur l'état de croyance initial de l'agent est optimisé ? Il est très important de souligner que nous souhaitons mesurer l'efficacité du critère du point de vue *objectif* d'un observateur extérieur au

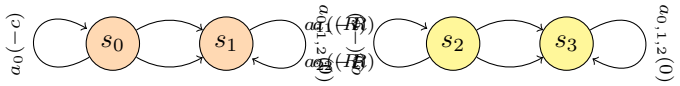

FIGURE 1 – Modèle de transition pour le POMDP (récompenses entre parenthèses)

système, qui connaîtrait parfaitement l'état de l'environnement à tout instant. Nous ne parlons pas ici des récompenses que l'agent croit ou espère gagner, mais des récompenses effectivement données à l'agent par l'observateur extérieur.

Nous pensons que la solution de ce problème dépend de la classe à laquelle la fonction J_λ appartient. Ainsi, même pour une classe simple du type $J_\lambda : J_\lambda(V, H) = (1 - \lambda)V + \lambda H, 0 \leqslant \lambda \leqslant 1\}$, nous ne pouvons pas trouver de solution algébrique générale. D'autres auteurs ont étudié des applications pour des valeurs fixées de λ avec projection immédiate (à un coup) dans le temps de l'entropie (Burgard *et al.*, 1997). D'autres ont formalisé des problèmes de perception active comme l'optimisation d'un POMDP basé sur la classe précédente de manière simplifiée mais sans étudier l'impact de λ dans l'accumulation des récompenses lors de l'exécution (Mihaylova *et al.*, 2002).

Dans la prochaine section, nous attirons l'attention sur l'importance de l'étude de nouveaux critères mixtes et non-linéaires en introduisant un exemple simple. Nous cherchons à démontrer l'impact des différentes valeurs de λ sur des récompenses effectivement gagnées *lors de l'exécution* de la politique, en partant d'un état de croyance donné. Dans les sections suivantes, nous définissons formellement deux critères : un additif et un multiplicatif qui peuvent être intéressants pour une optimisation plus adéquate de POMDP appliqués à la robotique. Finalement, et avant conclure, nous mettons en évidence l'importante de la considération des critères mixtes non linéaires dans un problème réaliste de reconnaissance et pistage multi-cible, lequel a été résolu avec un algorithme de l'état de l'art modifié pour nos nouveaux critères.

2 Exemple illustratif

Cette section propose d'étudier la différence de comportement obtenu *à l'exécution de la politique* avec une modification du critère classique optimisé pour un POMDP modélisant un problème jouet simple dont on peut calculer algébriquement une politique optimale. L'objectif est de montrer que le changement de critère *induit à une prudence de la part de l'agent par rapport à son état de croyance, en réduisant les possibilités d'erreurs pendant l'exécution de la politique*.

On définit un problème avec quatre états $\{s_0, s_1, s_2, s_3\}$ et deux observations $\{o_0, o_1\}$. Initialement, l'agent peut se retrouver en s_0 ou s_2, donc on a $b_0(s_0) = 1 - b_0(s_2)$, et o_0 (resp. o_1) correspondent à observer si l'agent se trouve en s_0 (resp. s_2). L'agent peut réaliser trois actions : a_0 est une action de perception qui coûte c et qui n'amène pas de changement d'état, alors que a_1 et a_2 mènent l'agent à des états absorbants comme montré dans la figure 1. En connaissant l'état réel s_0 ou s_2, les actions a_1 et a_2 donnent des récompenses opposées (R ou $-R$), ceci signifie que a_1 doit être choisi pour l'état réel s_0, et a_2 si l'agent se trouve en s_2. On note que $R > c > 0$. Intuitivement il y a deux "bonnes" politiques ici, qui dépendent directement de l'état de croyance initial de l'agent :

— éviter le coût de l'observation et choisir directement a_1 ou a_2 ;
— tout d'abord observer avec a_0 et en suite agir avec les actions a_1 ou a_2.

La matrice d'observation est définie par : $p(o|s') = \begin{bmatrix} 1 & 0.5 & 0 & 0.5 \\ 0 & 0.5 & 1 & 0.5 \end{bmatrix}$. Nous pouvons calculer les Q-valeurs sur $b(s)$, qui sont les valeurs pour chaque action supposant qu'une politique optimale est réalisée par la suite. Les Q-valeurs dépendent de $b_0(s_0)$ et $b_0(s_2)$:

$$
\begin{aligned}
Q^\pi(b, a_0) &= (R - c)(b_0(s_0) + b_0(s_2)) \\
Q^\pi(b, a_1) &= R(b_0(s_0) - b_0(s_2)) \\
Q^\pi(b, a_2) &= R(b_0(s_2) - b_0(s_0))
\end{aligned}
$$

Les Q-valeurs sur $b_0(s_0)$ sont montrées dans la figure 2, où la fonction de valeur est définie par $V^\pi(b) = \max_a Q(b, a)$. Nous voyons que la politique optimale dépend directement de l'état de croyance initial.

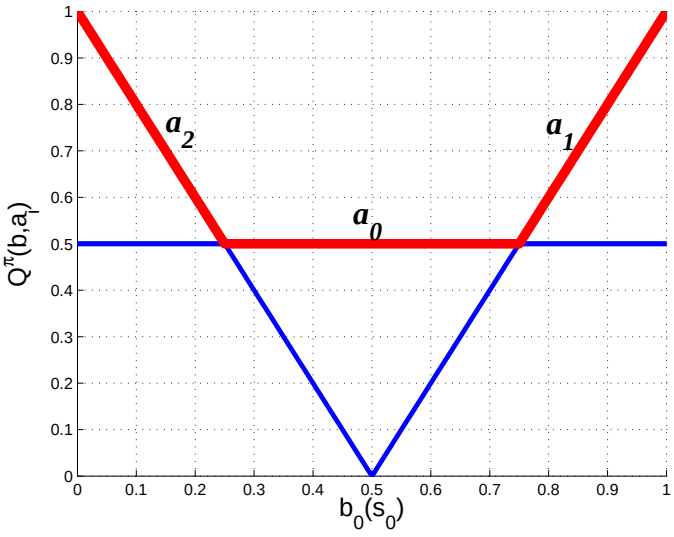

FIGURE 2 – Q-valeurs et fonction de valeur (en **gras**) sur $b_0(s_0)$

Modification du critère

Maintenant, nous ajoutons l'entropie de Shannon de l'état de croyance à chaque pas de temps à $b_t(s)$, c'est-à-dire que nous rajoutons l'espérance de la somme pondérée des entropies $H_t^\pi(b)$ notée par : $H_t^\pi = \sum_{t=0}^{N} H(b_t)$. Et donc, le nouveau critère peut se récrire comme : $J^\pi(b,\lambda) = (1-\lambda)V^\pi(b) + \lambda H^\pi(b)$.
La valeur de l'entropie de l'état de croyance $H(b_t)$ ne change pas lorsque la première politique est exécutée. D'autre part, quand la seconde politique est choisie, l'entropie est ramenée à zéro dès le second pas de temps. Après avoir choisi l'action a_1 or a_2, la valeur de l'entropie ne change pas. Donc, avec le critère mixte, la première politique est plus pénalisée que la deuxième, car le critère prend en compte la somme totale des entropies (en $t=0$ et $t=1$).

$$
\begin{aligned}
Q^\pi(b,a_0) &= (1-\lambda)(R-c)(b_0(s_0)+b_0(s_2)) + \lambda H(b_0)\\
Q^\pi(b,a_1) &= (1-\lambda)R(b_0(s_0)-b_0(s_2)) + \lambda(H(b_0)+H(b_1))\\
Q^\pi(b,a_2) &= (1-\lambda)R(b_0(s_2)-b_0(s_0)) + \lambda(H(b_0)+H(b_1))
\end{aligned}
$$

Afin d'illustrer le changement apporté au critère, nous avons calculé la meilleure politique pour différentes valeurs de λ. La figure 3 montre que la forme du critère mixte change beaucoup avec la variation de la valeur de λ de 0 à 1 : plus la valeur de λ est proche de 1, plus la première politique est pénalisée. Cette figure montre aussi que le critère n'est plus linéaire.
La figure 4 présente les récompenses réelles, effectivement obtenues par l'agent lors de l'exécution car celui-ci agit en fonction de son état de croyance $b(s)$, sans savoir qu'en réalité il se trouve initialement dans l'état s_0. Nous attirons l'attention sur la différence entre ces valeurs et la valeur supposée être obtenue dans la figure 2. Plus la valeur de λ est proche de 1, plus l'agent préfère observer plutôt qu'agir directement, et par conséquence il est moins pénalisé par son mauvais état de croyance initial (0.5 à la place de -1). Mais si son état de croyance est initialement correct, les récompenses recueillies diminuent aussi (0.5 à la place de 1). Donc, nous voulons établir un certain dégré de confiance sur $b_0(s)$ et donc trouver la valeur appropriée de λ pour le problème, c'est-à-dire rajouter une incertitude sur $b_0(s)$. Pour cet exemple il est possible de calculer une valeur de λ en fonction de $b_0(s_0)$, avec $b_(s_0) = b_0 s_0$ et en partant de $Q^\pi(b,a_0) = Q^\pi(b,a_2)$ et sachant que l'entropie $H(b_1) = H(b_0)$ pour l'action a_2 dans cet exemple.

$$
\begin{aligned}
\lambda &= \frac{2Rb_{s_0} - c}{2Rb_{s_0} - c + b_{s_0}\ln(b_{s_0}) + (1-b_{s_0}))\ln(1-b_{s_0}))}\\
&= \frac{2Rb_{s_0} - c}{2Rb_{s_0} - c + H(b_0)}
\end{aligned}
\qquad (1)
$$

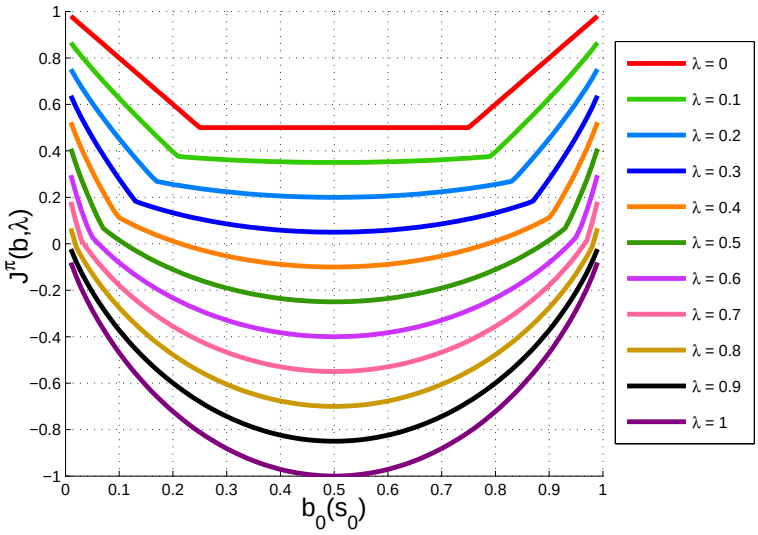

FIGURE 3 – Meilleur critère mixte basé sur l'état de croyance initial de l'agent pour différentes valeurs de λ : λ augmente de la courbe du haut vers la courbe du bas.

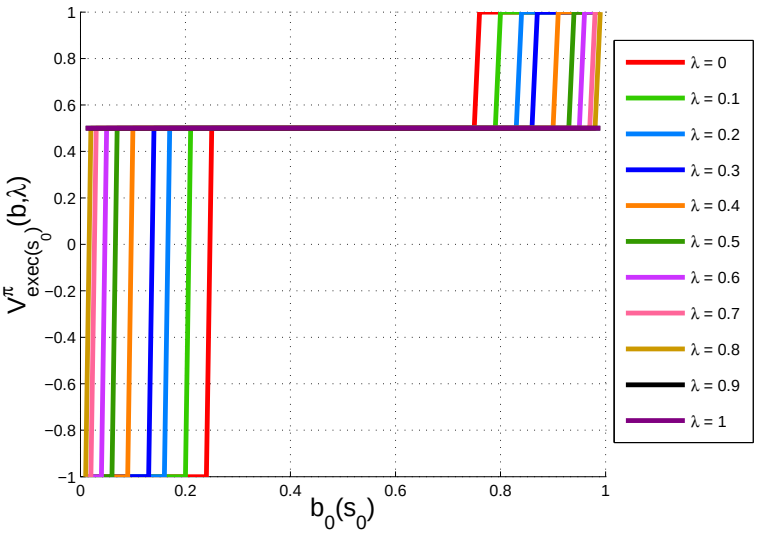

FIGURE 4 – Récompenses effectivement obtenues lors de l'exécution pour différentes valeurs de λ en fonction de l'état de croyance initial (qui détermine sa politique) ; λ augmente de l'intérieur vers l'extérieur, l'agent se trouve initialement dans l'état s_0.

Ce type de modification de critère peut être utile quand l'état de croyance de l'agent (*a priori*) $b_0(s)$ ne correspond pas à une distribution très proche de la réalité du système. Dans des cas réels, ce type d'erreur se présente assez souvent : le $b_0(s)$ utilisé dans le calcul de la politique peut ne pas être une bonne approximation de la réalité.

Contre exemple

Un contre exemple est détaillé dans la figure 5. Nous montrons qu'il n'y a pas de gain en rajoutant une mesure de l'entropie de l'état de croyance dans le critère s'il n'y a pas d'ambiguïté dans l'état d'arrivée. La fonction de valeur dépend seulement de l'état d'arrivée, et l'état d'où l'agent est initialement parti importe

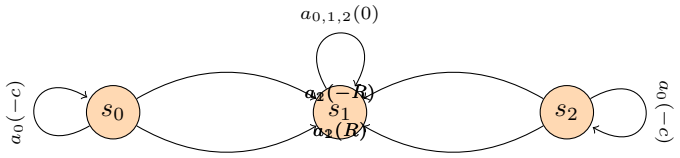

FIGURE 5 – Contre exemple pour l'addition de l'entropie dans le critère d'optimisation.

peu. Donc, la fonction de valeur va être pénalisée également par l'entropie pour chaque action.

$$Q^\pi(b, a_0) = R|b_0(s_1) - b_0(s_2)| - c$$
$$Q^\pi(b, a_1) = Q^\pi(b_0, a_2) \; = \; R|b_0(s_1) - b_0(s_2)|$$

Dans la section suivante, deux critères mixtes et non-linéaires pour les POMDP sont présentés et les modifications réalisées sur l'algorithme de l'état de l'art Symbolic-PERSEUS (Poupart, 2005) pour leur optimisation. Et, dans la section qui suit, un problème réaliste est résolu et les résultats obtenus confirment les intuitions soulevées avec cet exemple illustratif.

3 Critères d'optimisation hybrides pour les POMDP

3.1 Critère Additif

Notre première approche propose un critère d'optimisation additif : nous modélisons l'espérance de la somme pondérée des récompenses attribuées aux actions choisies, ajoutée à l'espérance de la somme pondérée des entropies des états de croyance stochastiques successifs, ceci à long terme. Ces deux valeurs sont elles-mêmes pondérées par une constante $\lambda \in [0; 1]$:

$$\begin{aligned} J^\pi(b) &= (1-\lambda)V^\pi(b) + \lambda H^\pi(b), \text{ avec} \\ V^\pi(b) &= E_\pi\left[\sum_{t=0}^{\infty} \gamma^t r(b_t, \pi(b_t)) \mid b_0 = b\right] \\ H^\pi(b) &= E_\pi\left[\sum_{t=0}^{\infty} \gamma^t H(b_t) \mid b_0 = b\right] \end{aligned} \tag{2}$$

Théorème: Équation de Bellman pour le critère additif 1
La fonction de valeur optimale du critère additif est la limite de la suite vectorielle définie par :

$$J_{n+1}(b) = \max_{a \in A}\left\{(1-\lambda).r(b,a) + \lambda.H(b) + \gamma \sum_{o \in \Omega} p(o|b,a) J_n(b_a^o(s'))\right\}$$

Proof. L'équation 3 peut se récrire :

$$J^\pi(b) = E_\pi\left[\sum_{t=0}^{\infty} \gamma^t \left((1-\lambda)r(b_t, \pi(b_t)) + \lambda H(b_t)\right) \mid b_0 = b\right] \tag{3}$$

ce qui montre que ce nouveau critère correspond au critère γ-pondéré classique dans lequel la récompense courante est ajoutée à l'entropie de croyance courante. Il s'agit donc d'un problème de maximisation γ-pondéré de récompenses artificielles, égales aux récompenses réelles ajoutées aux entropies. ■

Cette nouvelle équation de Bellman nous permet de calculer par programmation dynamique une politique qui pénalise la récompense immédiate par l'entropie de l'état de croyance dans les cas où la distribution sur les états est peu précise.

Heuristique

Symbolic-PERSEUS utilise une heuristique numérique pour déterminer l'ensemble initial des états de croyance pertinents. Il initialise la recherche des états de croyance atteignables par $V^{\pi}_{degrad} = \max_{s,a} r(s,a)$, calculés à partir d'un modèle dégradé, c'est-à-dire avec une heuristique *admissible* dont la valeur doit être plus petite que la valeur optimale. La définition de notre nouveau critère nécessite donc de modifier aussi cette heuristique, afin que celle-ci prenne en compte une valeur minimale de H dans l'initialisation du calcul des états atteignables pour le nouveau critère J^{π}, c'est-à-dire un J_0.

Théorème: Heuristique pour le critère additif 1
Une heuristique admissible pour le critère additif est donnée par :

$$J_0 = \frac{(1-\lambda)V^{\pi}_{degrad} - \lambda \log_{10}(n)}{1-\gamma} \tag{4}$$

Proof. Par optimisation lagrangienne, la valeur minimale de $H(b)$ sous la contrainte $\sum_{i=1}^{n} b(s_i) = 1$ est :

$$H(b)_{min} = n\frac{1}{n}\log_{10}\left(\frac{1}{n}\right) = -\log_{10}(n) \tag{5}$$

$$\text{Ainsi : } J^{\pi} \geq \frac{(1-\lambda)V^{\pi}_{degrad} - \lambda \log_{10}(n)}{(1-\gamma)} = J_0 \tag{6}$$

■

Discussion

Les critères présentés ici ne sont pas linéaires par morceaux, mais des algorithmes tels que PBVI (Pineau *et al.*, 2003), HSVI (Smith & Simmons, 2004), PERSEUS (Spaan & Vlassis, 2004) et Symbolic-PERSEUS (Poupart, 2005), qui approximent le critère par génération stochastique d'états de croyances locaux, peuvent approcher ces critères non-linéaires, sachant que toute fonction continue est approchable par une fonction linéaire par morceaux.

3.2 Critère Multiplicatif

Une deuxième approche propose un critère d'optimisation multiplicatif : nous modélisons l'influence de l'espérance de la somme pondérée des récompenses des actions, pondérées par l'inverse de la valeur absolue de l'entropie *immédiate* de l'état de croyance. Plus l'entropie de l'état de croyance est petite, plus le critère est grand, ce qui va dans le sens d'un gain d'information explicite.

$$J^{\pi}(b) \quad = \quad E_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t \frac{r(b_t, \pi(b_t))}{|H(b_t)|} \mid b_0 = b\right] \tag{7}$$

Théorème: Équation de Bellman pour le critère multiplicatif 1
La fonction de valeur optimale du critère multiplicatif est la limite de la suite vectorielle définie par :

$$J_{n+1}(b) = \max_{a \in A}\left\{\frac{r(b,a)}{|H(b)|} + \gamma \sum_{o \in \Omega} p(o|b,a) J_n(b^o_a(s'))\right\} \tag{8}$$

Proof. Ce nouveau critère correspond au critère γ-pondéré classique dans lequel la récompense courante est divisée par l'entropie de croyance courante. Il s'agit donc d'un problème de maximisation γ-pondéré de récompenses artificielles, égales aux récompenses réelles divisées par les entropies. ■

Cette nouvelle équation de Bellman nous permet de calculer par programmation dynamique une politique qui pondère la récompense immédiate par l'inverse de la valeur absolue de l'entropie de l'état de croyance.

Cibles	A	B
cible 1	0.2 (×)	0.8
cible 2	0.8	0.2 (×)

TABLE 1 – État de croyance initial de l'agent sur les cibles ; le signe (×) indique la nature réelle des cibles.

Heuristic

Comme dans le cas additif, les heuristiques utilisées pour l'initialisation du calcul des états de croyance atteignables, ainsi que pour le calcul de la politique ont dû être modifiées pour prendre en compte la modification du critère.

Théorème: Heuristique pour le critère multiplicatif 1
Une heuristique admissible pour le critère multiplicatif est donnée par :

$$J_0 = \frac{V^{\pi}_{degrad}}{(|-\log_{10}(n)|)(1-\gamma)} \tag{9}$$

4 Exemple robotique

Scénario étudié

Le modèle étudié traite d'un hélicoptère autonome qui cherche à identifier et pister deux cibles mobiles. Ces cibles sont de natures différentes, l'une du type *A* et l'autre du type *B*. Le but de l'hélicoptère autonome est de se poser sur la cible *A*, sans connaître initialement la nature des cibles. Ce scénario mêle à la fois des objectifs de mission et de perception. Il est intéressant pour nous car l'optimisation des récompenses implique (implicitement) la diminution de l'entropie de croyance : il est en effet nécessaire de réduire l'incertitude sur la nature des cibles afin de réaliser la mission.

Initialement, l'hélicoptère autonome dispose d'une connaissance *a priori* des cibles. Il doit, à partir de ses actions, pister et identifier chaque cible afin d'accomplir sont but final. Nous soulignons que, pour les simulations étudiées dans ce travail, nous avons donné un état de croyance initial inversé (mauvaise connaissance *a priori*) par rapport aux natures réelles des cibles. Ces valeurs sont montrées dans le tableau 1. La cible 1 a pour nature réelle le type *A* et la cible 2, le type *B*.

L'espace de déplacement de l'hélicoptère est modélisé par une grille $3 \times 3 \times 3$, et celui des cibles par une grille $3 \times 3 \times 1$, les cibles évoluant au sol uniquement. L'hélicoptère peut réaliser 7 actions : avancer en x, avancer en y, avancer en z (monter), reculer en x, reculer en y, reculer en z (descendre), et atterrir. Il ne

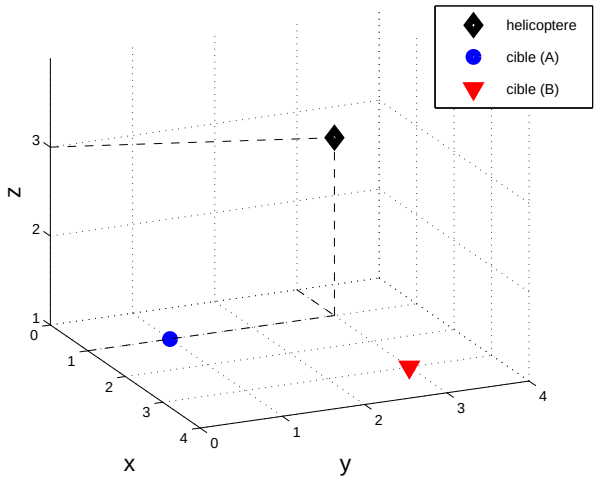

FIGURE 6 – Position initial de l'hélicoptère, cible 1 (A) et cible 2 (B).

peut pas réaliser plus d'une action à la fois. Les actions de déplacement de l'hélicoptère ont une probabilité d'échec de 10%, sauf l'action $atterrir$ qui est déterministe.

La position de chaque cible est complètement observable par l'hélicoptère. Cependant, les cibles changent de place en x et/ou en y 1 fois sur 20 déplacements de l'hélicoptère, mais celui-ci ne peut pas prévoir cette évolution. L'atterrissage est autorisé pour l'hélicoptère seulement s'il est au-dessus d'une cible et à une altitude de $z = 2$. Une fois que l'hélicoptère a atterri (sur la cible *A* ou *B*) , il ne peut plus décoller.

La récompense a été modélisée par un coût pour chaque action dépendant de l'état d'arrivée. Chaque action de déplacement en x, y et z a un coût de 1 ; l'action $atterrir$ implique un coût ou une récompense de 100 suivant la cible, le but étant d'atterrir sur la cible *A*. Le modèle d'observation de la nature des cibles dépend de la distance euclidienne entre l'hélicoptère et les cibles :

$$\begin{aligned} p(o' = \mathrm{A} \mid s' = \mathrm{A}) = p(o' = \mathrm{B} \mid s' = \mathrm{B}) &= \frac{1}{2}\left(e^{\frac{-d}{D}} + 1\right) \\ p(o' = \mathrm{A} \mid s' = \mathrm{B}) = p(o' = \mathrm{B} \mid s' = \mathrm{A}) &= \frac{1}{2}\left(1 - e^{\frac{-d}{D}}\right) \end{aligned}$$

où d est la distance géométrique entre l'hélicoptère et la cible, et D un facteur de réglage de la descente de l'exponentielle. Cette fonction d'observation nous permet de modéliser le gain d'information lorsque l'hélicoptère s'approche de la cible : plus l'hélicoptère est proche de la cible observée, plus la probabilité qu'il observe la nature réelle de la cible est grande. L'hélicoptère observe de manière déterministe les autres variables d'état, comme par exemple sa position et celle des cibles. Remarquons que ce modèle n'a pas pour but de tester l'efficacité d'un algorithme en termes de temps de calcul ou de mémoire utilisée, mais de comparer différents critères d'optimisation pour un même problème avec un critère mixte.

4.1 Protocole expérimental

Les critères d'optimisation des POMDP sont fonctions de l'état de croyance de l'agent, car ils sont optimisés du point de vue *subjectif* de l'agent, qui n'a pas accès directement à l'état de l'environnement. Or, le critère de performance que nous souhaitons mesurer est basé sur les récompenses qui seront *réellement* cumulées lors de l'exécution de la politique optimisée, moyennant l'incertitude sur l'effet des actions uniquement. Autrement dit, nous souhaitons mesurer l'efficacité de chaque critère du point de vue *objectif* d'un observateur extérieur au système, qui connaîtrait parfaitement l'état de l'environnement à tout instant. Dans cet article, cet observateur omniscient sera un simulateur des politiques optimisées.

Pour chaque politique optimisée, nous réalisons 100 simulations sur un horizon de 50 actions successives. Pour $\gamma = 0.9$ cet horizon est considéré suffisamment grand pour obtenir une bonne approximation des critères en horizon infini. La fonction de valeur objective, qui, elle, dépend de l'état réel courant de l'environnement, est calculée suivant l'équation 10.

$$V^{\pi}(s_t) = E^{\pi}_{100\, simulations}\left[\sum_{k=t}^{50} \gamma^k r^{\pi}(s_k) \mid s_t\right] \tag{10}$$

Nous comparons les politiques optimisées avec les différents critères sur la base de cette même valeur objective, qui, quelque soit le critère d'optimisation, sera toujours la valeur réellement gagnée par l'agent autonome lors de l'exécution de la politique.

L'intérêt de l'analyse de $V^{\pi}(s)$ à la place de $V^{\pi}(b)$ est de vérifier le nombre de fois où l'hélicoptère a accompli correctement sa mission en partant d'un état de croyance inversé. $V^{\pi}(s)$ a été calculé par le simulateur qui, lui, connaît l'état réel des cibles. $V^{\pi}(s)$ permet aussi de comparer le pessimisme du critère classique par rapport au critère pondéré par le gain d'information.

Afin d'étudier la vitesse de convergence de l'entropie de croyance de l'agent, nous calculons également la moyenne statistique de l'entropie de croyance courante, comme indiqué dans l'équation 11.

$$H^{\pi}(b_t) = E^{\pi}_{100\, simulations}\left[\sum_{k=t}^{50} \gamma^k H^{\pi}(b_k) \mid b_t\right] \tag{11}$$

Notons que cette mesure est subjective et propre à l'agent, contrairement à la mesure précédente qui est objective et propre au simulateur.

4.2 Résultat des simulations

4.2.1 Critère Additif

Des politiques ont été calculées pour différentes valeurs du coefficient λ : 0, 0.5 et 1. Le cas 0 est identique au critère γ-pondéré classique. On rappelle que ce critère cherche à optimiser uniquement l'espérance de la récompense pondérée attribuée aux actions et à l'accomplissement de la mission. Le deuxième cas cherche à donner la même importance à l'accomplissement de la mission et à l'acquisition explicite d'information, le troisième optimise le gain d'information.

Dans la figure 7, la moyenne des fonctions de valeur $V^{\pi}(s)$ (équation 10) est montrée pour les trois cas. Dans le premier cas, $\lambda = 0$, la fonction de valeur statistique $V^{\pi}(s)$ part d'une valeur négative, ce qui s'explique par la façon dont elle est calculée. Les atterrissages sur la bonne cible (réalisés au bout de 10 pas de simulation ou plus) comptent moins que ceux sur la mauvaise (réalisés au bout de 3 ou 4 pas de simulation) dans le calcul de $V^{\pi}(s_0)$ à cause de la pondération γ. Nous pensons que les atterrissages sur la mauvaise cible sont probablement dus à la petite taille de la grille, empêchant l'hélicoptère autonome d'acquérir plus d'information avant de se poser. L'hélicoptère, qui part d'un état de croyance inversé, tend à atterrir sur la cible *B* au bout de 3 ou 4 pas de simulation, car il croit à ce moment que cette cible est la cible *A*. Par contre, pour les simulations où l'hélicoptère a pu acquérir plus d'information de son environnement, il est montré qu'en moyenne, l'hélicoptère autonome inverse son état de croyance et atterrit sur la bonne cible. D'où l'inversion observée de la courbe de valeur, qui montre que l'agent réagit bien au pire cas, avec les deux critères.

Pour le deuxième cas, $\lambda = 0.5$, nous vérifions que la valeur du critère part cette fois-ci d'une valeur positive : les atterrissages sur la bonne cible, réalisés plus tôt maintenant que pour le critère classique, comptent plus dans le calcul de $V^{\pi}(s_0)$ à cause de la pondération γ. Notre contribution est ici montrée, puisque l'agent cherche maintenant de façon explicite à acquérir plus d'information de son environnement, ce qui lui permet d'inverser plus tôt son état de croyance et finalement de se poser plus fréquemment sur la bonne cible. Un des problèmes du critère classique est justement sa linéarité en fonction de $b(s)$: il considère par exemple qu'un état de croyance avec 60% de chances "vaut" 60% de cette récompense. La non-linéarité de ce critère permet d'évaluer plus finement la valeur de $b(s)$, en donnant plus de poids aux plus grandes certitudes. La figure 7 montre bien que l'agent a réellement cumulé plus de récompenses avec notre critère additif qu'avec le critère γ-pondéré classique. Cela nous permet de conclure que l'addition de l'influence de l'entropie de croyance dans le critère d'optimisation pousse l'agent autonome à mieux percevoir son environnement pour ensuite accomplir mieux sa mission lors de l'exécution de la politique, ce qui est l'objectif généralement visé par les concepteurs d'un système autonome.

Pour le troisième cas, $\lambda = 1$, le critère optimise le gain d'information uniquement. La figure 7 montre que la moyenne de $V^{\pi}(s)$ reste proche de zéro : l'hélicoptère ne cherche pas à atterrir, car le gain d'atterrissage modélisé dans les récompenses n'est pas pris en compte.

La figure 8 compare le critère subjectif $H^{\pi}(b)$, équation 11, pour les 3 cas étudiés. On vérifie bien que pour le deuxième cas, la moyenne de la somme de l'entropie converge plus rapidement que pour le critère classique. En effet, l'hélicoptère a besoin d'observer son environnement avant d'accomplir sa mission, ce qui le contraint à réduire l'incertitude sur son état de croyance plus rapidement. La figure 8 montre aussi que l'optimisation de $V^{\pi}(b)$ permet de réduire l'incertitude plus vite que celle de l'optimisation de $H^{\pi}(b)$ uniquement, car il faut implicitement réduire l'incertitude pour atterrir sur la bonne cible. De plus, ce travail montre également que l'optimisation de $H^{\pi}(b)$ n'optimise pas nécessairement sa vitesse de décroissance durant l'exécution de la stratégie. Ici, les buts de la mission, intégrés au critère additif, poussent à réduire l'incertitude plus vite qu'en optimisant uniquement $H^{\pi}(b)$.

Nous pouvons donc conclure que notre critère additif est aussi optimal *en terme de gain d'information* que le critère purement entropique, et qu'il converge plus vite lors de l'exécution que les deux critères extrêmes (purement entropique ou γ-pondéré classique). Ceci est la cause d'un deuxième constat fondamental : le critère additif, en forçant la politique optimisée à sacrifier de temps en temps des actions de gain de récompenses pour des actions de gain d'information, cumule en réalité plus de récompenses lors de l'exécution que le critère γ-pondéré classique.

4.2.2 Critère Multiplicatif

La figure 9 compare la statistique des récompenses cumulées lors des simulations pour le critère γ-pondéré et notre critère multiplicatif. La courbe du critère multiplicatif reste toujours autour de zéro, ce qui

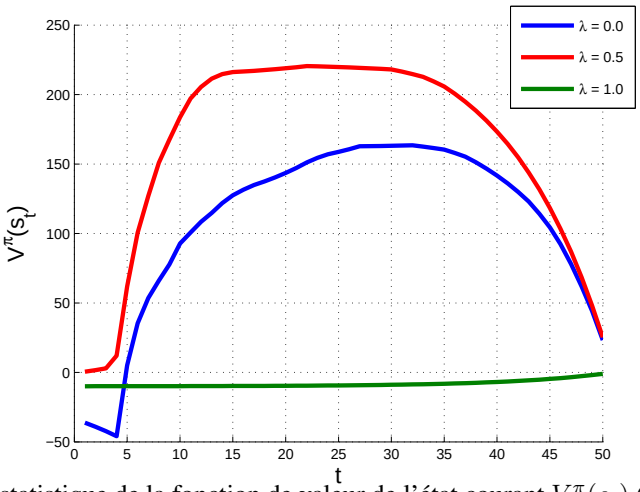

FIGURE 7 – Moyenne statistique de la fonction de valeur de l'état courant $V^{\pi}(s_t)$ (l'état change le long de la courbe).

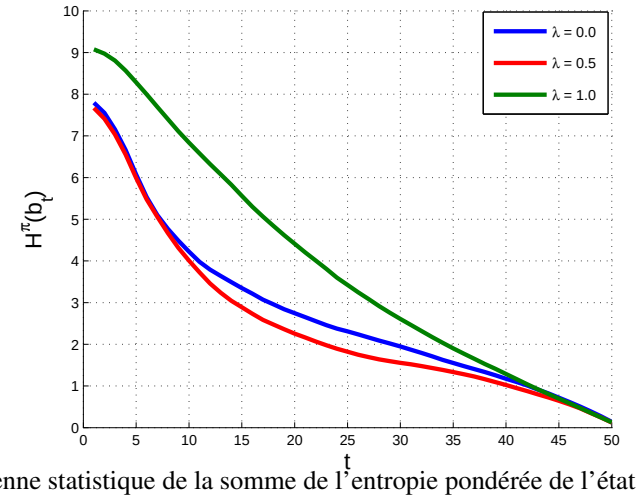

FIGURE 8 – Moyenne statistique de la somme de l'entropie pondérée de l'état de croyance $H^{\pi}(b_t)$.

montre que l'hélicoptère ne cherche pas à atterrir. Ainsi, ces résultats nous permettent de dire que le critère multiplicatif donne plus d'importance à la diminution de l'entropie de l'état de croyance qu'au but de la mission.

Le changement du critère à optimiser n'est pas un changement linéaire comme l'est le critère additif, et donc, la récompense due à l'accomplissement de la mission joue comme une pondération qui force l'hélicoptère autonome à s'approcher de l'une ou l'autre cible afin de valider sa nature. Ceci est vérifié sur la figure 10 : la vitesse de convergence de $H^{\pi}(b)$ du critère multiplicatif est plus importante que celle du critère classique. Par contre, l'accomplissement de la mission n'est plus le paramètre pris en compte comme but principal.

5 Conclusion et perspectives

Dans cet article, nous avons proposé deux nouveaux critères d'optimisation mixtes pour les POMDP, l'un multiplicatif et l'autre additif, qui agrègent le gain cumulé d'information (perception) et le gain cumulé de récompenses (mission), pondérés et moyennés sur un horizon infini. Nous avons mis à jour et prouvé l'optimalité des équations de Bellman pour ces nouveaux critères. Nous avons également proposé de nouvelles heuristiques admissibles pour ces critères, afin qu'ils puissent être utilisés dans des algorithmes heuristiques comme Symbolic-Perseus.

Nous avons montré expérimentalement que ces deux critères permettent à l'agent autonome d'acquérir

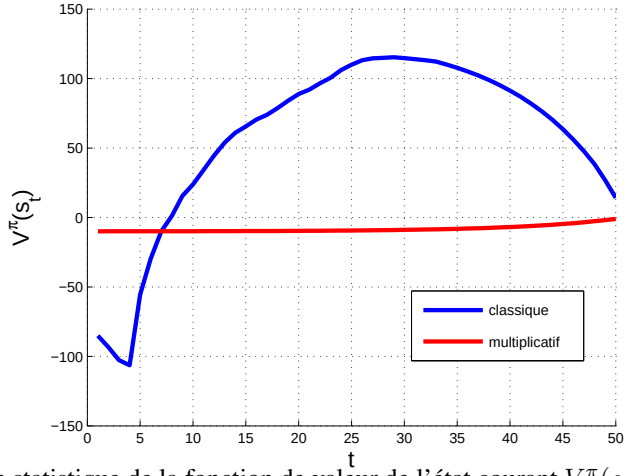

FIGURE 9 – Moyenne statistique de la fonction de valeur de l'état courant $V^\pi(s_t)$ (l'état change le long de la courbe).

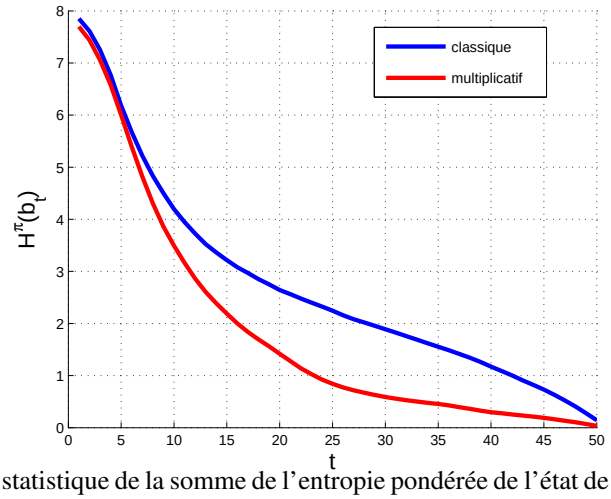

FIGURE 10 – Moyenne statistique de la somme de l'entropie pondérée de l'état de croyance courant $H^\pi(b_t)$.

plus rapidement de l'information sur son environnement, et donc d'estimer plus vite son état réel, en comparaison à des critères classiques qui prennent uniquement en compte soit le gain d'informations, soit le gain de récompenses. De plus, grâce au critère additif, l'agent cumule en réalité plus de récompenses lors de l'exécution de la politique optimisée, par rapport au critère γ-pondéré classique qui ne prend pas en compte le gain d'informations explicite. En quelque sorte, la prise en compte explicite de l'entropie de croyance conjointement aux récompenses pousse l'agent autonome à acquérir de l'information pour débiaiser sa vue subjective des récompenses qu'il croît pouvoir cumuler mais qu'il ne cumulera peut-être pas en raison de sa connaissance imparfaite de l'environnement.

Dans le futur, nous pensons étudier plus finement l'influence du coefficient de pondération λ dans le critère additif. Nous pensons qu'il existe un coefficient λ optimal qui dépend de la classe du modèle étudié et du degré de confiance donné à l'état de croyance initial. Ce coefficient λ permettra de maximiser les récompenses obtenues lors de l'exécution de la politique en tenant compte d'une incertitude sur b_0. Nous comptons proposer un algorithme qui optimise en même temps le coefficient λ et la politique pour une fonction de valeur qui peut être non linéaire, par rapport à la classe du problème modélisé.

Références

BURGARD W., FOX D. & THRUN S. (1997). Active mobile robot localization. In *Proceedings of IJCAI-97* : Morgan Kaufmann.

CASSANDRA A., KAELBLING L. & KURIEN J. (1996). Acting under uncertainty : Discrete Bayesian models for mobile-robot navigation. In *In Proceedings of IEEE/RSJ*.

DEINZER F., DENZLER J. & NIEMANN H. (2003). Viewpoint selection-planning optimal sequences of views for object recognition. *Lecture notes in computer science*, p. 65–73.

EIDENBERGER R., GRUNDMANN T., FEITEN W. & ZOELLNER R. (2008). Fast parametric viewpoint estimation for active object detection. In *Proceeding of the IEEE International Conference on Multisensor of Fusion and Integration for Intelligent Systems (MFI 2008), Seoul, Korea*.

KAELBLING L., LITTMAN M. & CASSANDRA A. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, **101**(1-2), 99–134.

MIHAYLOVA L., LEFEBVRE T., BRUYNINCKX H., GADEYNE K. & SCHUTTER J. D. (2002). Active sensing for robotics – a survey. In *5th Intl Conf. On Numerical Methods and Applications*, p. 316–324.

PALETTA L. & PINZ A. (2000). Active object recognition by view integration and reinforcement learning. *Robotics and Autonomous Systems*, **31**, 71–86.

PINEAU J., GORDON G. & THRUN S. (2003). Point-based value iteration : An anytime algorithm for POMDPs. In *Proc. of IJCAI*.

POUPART P. (2005). *Exploiting structure to efficiently solve large scale partially observable Markov decision processes*. PhD thesis, University of Toronto.

SMITH T. & SIMMONS R. (2004). Heuristic search value iteration for POMDPs. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, p. 520–527 : AUAI Press.

SONDIK E. & LABS S. U. C. S. E. (1971). *The Optimal Control of Partially Observable Markov Processes.* Stanford University, California.

SPAAN M. & VLASSIS N. (2004). A point-based POMDP algorithm for robot planning. In *IEEE International Conference on Robotics and Automation*, volume 3, p. 2399–2404 : IEEE ; 1999.

SPAAN M. & VLASSIS N. (2005). Perseus : Randomized point-based value iteration for POMDPs. *JAIR*, **24**, 195–220.

SRIDHARAN M., WYATT J. & DEARDEN R. (2008). HiPPo : Hierarchical POMDPs for Planning Information Processing and Sensing Actions on a Robot. In *Proc. of ICAPS*.