Optimisation des Processus Décisionnels de Markov Partiellement Observables avec prise en compte explicite du gain d'information

Caroline Ponzoni Carvalho [1,2] Florent Teichteil-Königsbuch [2] Guillaume Infantes [2]

[1] ISAE - Institut Supérieur de l'Aéronautique et de l'Espace
10 avenue Edouard Belin - BP 54032 - 31055 TOULOUSE Cedex 4
[2] ONERA - Office National d'Étude et Recherche Aérospatiale
2 avenue Edouard Belin - BP 74025 - 31055 TOULOUSE Cedex 4
{caroline.carvalho,florent.teichteil,guillaume.infantes}@onera.fr

Résumé

Traditionnellement, les travaux de recherche en décision séquentielle dans l'incertain avec observabilité partielle reposent sur les Processus Décisionnels de Markov Partiellement Observables (POMDP), optimisés avec un critère de maximisation de revenus cumulés pondérés sur un horizon d'action donné. Or, ce critère est pessimiste dans la mesure où la décision est optimisée sur une distribution de probabilité sur l'état de croyance de l'agent autonome, sans que l'algorithme ne réduise explicitement cette incertitude. Autrement dit, les critères classiques d'optimisation des POMDP raisonnent sur toutes les hypothèses possibles, sans favoriser explicitement les actions qui pourraient acquérir de l'information et réduire le champ d'hypothèses. Au contraire, les travaux en traitement d'image et particulièrement en perception active s'intéressent plutôt à trouver les actions qui minimisent l'entropie de croyance, c'est-à-dire l'incertitude sur l'état caché, mais sans optimiser une récompense globale liée à la mission du robot. Ainsi, afin de résoudre au mieux des problèmes robotiques alliant à la fois des objectifs de perception et de mission, nous proposons deux nouveaux critères mixtes, l'un additif et l'autre multiplicatif, qui agrègent les récompenses cumulées (mission) et les entropies de croyance cumulées (perception), toutes deux pondérées sur un horizon d'action commun. À l'aide d'évaluations statistiques sur plusieurs exécutions de la politique optimisée, nous montrons que nos critères mixtes sont optimaux par rapport à un critère purement entropique, et que le critère additif améliore même un critère basé purement sur les récompenses de la mission. Ce dernier point démontre que le critère classique, qui repose uniquement sur les récompenses cumulées, n'est pas optimal lors de l'exécution, car il ne prend pas en compte explicitement le gain d'information et la réduction de l'incertitude sur l'état caché du système.

Mots Clef

POMDP, perception active, entropie de croyance, critère d'optimisation, décision séquentielle dans l'incertain.

Abstract

Research on sequential decision making under uncertainty with partial observability often relies on Partially Observable Markov Decision Processes (POMDPs) whose optimization criterion is a sum of discounted cumulated rewards over a given horizon. Yet, because of partial observability, this criterion optimizes action choices on a continuous set of hypotheses without aiming at explicitly reducing the field of possible hypotheses. On the contrary, works on active perception rather consist in finding actions that minimize the belief entropy, which is a probability distribution over hypotheses, but without trying to maximize a global reward representing the robot's mission. Therefore, in order to solve robotics problems with both perception and mission goals, we propose two new optimization criteria for POMDPs, multiplicative and additive, that aggregate the cumulated rewards (mission) and cumulated entropies (perception), both discounted over a common horizon. Based on statistical evaluations of policies optimized with each criterion, we demonstrate that our mixed criteria are optimal compared with the purely entropic criterion, and that the additive one even improves at run time the traditional γ-discounted criterion. It proves that the latter criterion, which only relies on cumulated rewards, is not optimal at run time because it does not explicitly take into account the information gathered by executing the policy.

Keywords

POMDP, active perception, belief's entropy, optimization criteria, sequential decision-making under uncertainty.

1 Introduction

Longtemps délaissés par les roboticiens en raison de leur complexité algorithmique, les Processus Décisionnels de Markov Partiellement Observables (POMDP) bénéficient aujourd'hui d'algorithmes d'optimisation efficaces permettant de résoudre de manière approchée des problèmes de très grande taille. Les POMDP sont un cadre formel pour la décision séquentielle en présence d'incertitudes sur les

effets des actions et sur les observations. Ce deuxième type d'incertitude est de loin prépondérant dans la complexité des algorithmes de résolution des POMDP, car il nécessite de raisonner sur une distribution de probabilité sur les états du système autonome, autrement dit sur un espace continu (de support infini). D'un autre côté, l'incertitude sur les observations apporte une richesse de modélisation qui démarque le modèle POMDP des autres modèles de planification, aussi bien déterministes que prenant en compte uniquement l'incertitude sur les effets des actions : en effet, l'incertitude sur les observations permet de formuler des hypothèses sur l'état réel de l'environnement, qui pourront être corroborées ou infirmées au fur et à mesure que de nouvelles observations sont collectées.

Ainsi, la mise à jour de l'ensemble des hypothèses possibles, appelé *état de croyance*, en fonction des observations passées est une problématique centrale des algorithmes d'optimisation des POMDP. Traditionnellement, un problème POMDP consiste à maximiser un critère numérique qui dépend de récompenses associées aux transitions entre les états. Le critère γ-pondéré, très utilisé en robotique, repose sur la moyenne de la somme pondérée des récompenses cumulées en partant de tout état de croyance de l'agent autonome. Intuitivement, la maximisation du critère γ-pondéré devrait réduire l'ensemble des hypothèses possibles, puisque les récompenses sont associées aux transitions entre états et non entre états de croyance : maximiser les récompenses cumulées nécessite *implicitement* de connaître au mieux l'état courant du système, c'est-à-dire de réduire *implicitement* l'incertitude sur l'état courant.

D'autre part, le domaine de recherche de la perception active s'intéresse à contrôler un ensemble de moyens d'observation (par exemple, des capteurs) afin de percevoir au mieux l'environnement dans un ensemble d'images. Là aussi, les différents modèles utilisés se basent sur un ensemble d'hypothèses initiales, généralement probabilistes, qui sont mises à jour au fur et à mesure que des actions d'observation sont effectuées. Le but d'un algorithme de perception active est de choisir les actions qui réduisent au plus vite le nombre d'hypothèses possibles. Certains travaux reposent sur un modèle de type POMDP, mais sans récompenses et dont le critère d'optimisation repose sur la notion d'*entropie*, qui est une mesure de la quantité d'information perçue : plus l'entropie est grande, plus le nombre d'hypothèses est grand. En termes probabilistes, une équiprobabilité sur les hypothèses maximise l'entropie de l'état de croyance. Contrairement à l'approche mentionnée dans le paragraphe précédent, cette approche consiste à réduire *explicitement* l'incertitude sur l'état courant du système.

Dans cet article, nous proposons d'unifier les deux approches précédentes au travers de deux nouveaux critères d'optimisation, l'un additif et l'autre multiplicatif, basés à la fois sur la moyenne pondérée des récompenses cumulées et sur l'entropie de croyance. L'objectif de cette étude est double :

– proposer un modèle unifié pour résoudre des problèmes robotiques qui allient à la fois des objectifs de perception (réduction de l'entropie de croyance) et des objectifs de mission (maximisation de la moyenne pondérée des récompenses cumulées) ;
– étudier l'apport d'une réduction explicite des hypothèses courantes sur la maximisation de la moyenne pondérée des récompenses effectivement cumulées lors de plusieurs simulations de la stratégie d'action optimisée.

Ce second point peut paraître surprenant, puisque la maximisation du critère γ-pondéré sans prise en compte de l'entropie de croyance produit effectivement une stratégie optimale au regard de ce critère. Néanmoins, ce critère est maximisé sur la base de l'état de croyance, et non sur celle de l'état réel du système, inconnu de l'algorithme et de l'agent autonome. Or, il est possible de simuler sur un grand nombre d'expériences la stratégie optimisée, et d'analyser les récompenses réellement cumulées lors de ces expériences successives. La moyenne statistique de ces récompenses cumulées, différentes du critère γ-pondéré, nous permet d'évaluer la performance *objective* — vue par un observateur omniscient qui connaîtrait parfaitement l'état du système — de la stratégie d'action utilisée. De fait, c'est cette métrique de performance qui intéresse réellement le concepteur du système autonome, bien qu'on ne puisse pas l'optimiser directement. Nous montrerons sur un scénario de reconnaissance et pistage de cibles mobiles par un hélicoptère autonome[1], que notre nouveau critère d'optimisation additif, qui prend explicitement en compte l'entropie de croyance, améliore sensiblement cette métrique de performance, par rapport au critère d'optimisation traditionnel basé uniquement sur la moyenne pondérée des récompenses cumulées.

Cet article est organisé de la manière suivante : dans la section 2, nous présentons le modèle général des POMDP ainsi que les critères d'optimisation les plus utilisés, en décision dans l'incertain comme en perception active. Dans la section 3, nous proposons deux nouveaux critères d'optimisation pour les POMDPs, qui agrègent la moyenne pondérée des récompenses cumulées et le gain d'information acquis. La section 4 présente le scénario d'étude utilisé, qui allie des objectifs de perception et de mission, et les résultats obtenus. Enfin, dans la section 5, nous concluons cet article par un bilan de nos contributions et des commentaires sur des travaux futurs.

2 Contexte et travaux connexes

2.1 POMDP

Les Processus Décisionnels Markoviens Partiellement Observables (POMDP) sont plusieurs problèmes de décision en séquence qu'un agent doit résoudre en présence d'incertitudes sur les effets des actions et l'état réel de l'environnement, où chaque décision courante influence la résolution du problème qui suit [5, 10]. Résoudre un POMDP revient

[1] Ce scénario multidisciplinaire est étudié dans le projet SPIDER (Systèmes de Perception et d'Interprétation Dynamiques Embarqués pour l'environnement uRbain) à l'ONERA.

à contrôler l'agent pour qu'il se comporte de manière optimale à long terme dans un environnement incertain. Dans le cas général, l'agent n'a accès qu'à des informations partielles sur le processus à contrôler : il ne connaît pas l'état réel du processus et ne peut accéder qu'à une observation partielle sur cet état.

Un POMDP est défini par un octuple (S, A, T, Ω, O, r, b_0, t), où S représente l'espace d'état, A l'espace d'actions, T la fonction des probabilités de transition entre états $p(s_{t+1} \mid s_t, a_t)$, Ω l'espace des observations, O la fonction des probabilités d'observation sur les états $p(o_{t+1} \mid s_{t+1})$, associé à Ω ; $r(s_t, a_t)$ la fonction de récompense sur les transitions entre états, b_0 la distribution de probabilité initiale sur les états, ou état de croyance initial (t est l'axe temporel). Comme l'agent n'a pas accès à l'état du système, il maintient et met à jour une distribution de probabilité sur les états, c'est-à-dire un état de croyance défini par b_t, sur la base de l'historique des observations. L'état de croyance doit être mis à jour à chaque action réalisée et observation reçue. Cette mise à jour repose sur la règle de Bayes [5] :

$$\begin{aligned} b_o^a(s') &= p(s' \mid b, a, o) \\ &= \frac{p(o \mid s', a) \sum_{s \in S} p(s' \mid s, a) b(s)}{\sum_{s' \in S} p(o \mid s', a) \sum_{s \in S} p(s' \mid s, a) b(s)} \end{aligned} \quad (1)$$

Résoudre un POMDP consiste généralement à maximiser l'espérance du revenu accumulé par l'agent autonome à long terme, c'est-à-dire, en projetant dans un futur suffisamment lointain, l'influence d'une action réalisée à l'instant présent sur le revenu total pondéré cumulé, en supposant que cette action est optimale (Π est l'ensemble des stratégies d'action possibles) :

$$V^*(b) = \max_{\pi \in \Pi} E_\pi \left[\sum_{t=0}^{\infty} \gamma^t r(b_t, \pi(b_t)) \mid b_0 = b \right] \quad (2)$$

Une solution optimale d'un POMDP est dans ce cas une politique π^* déterministe qui associe à chaque état de croyance b une action optimale $a \in A$. La valeur d'une politique optimale π^* est égale à la fonction de valeur optimale V^*, qui satisfait l'équation d'optimalité de Bellman, $V^* = \mathcal{L}V^*$:

$$V^*(b) = \max_{a \in A} \left\{ r(b, a) + \gamma \sum_{o \in \Omega} p(o \mid a, b) V^*(b_a^o) \right\} \quad (3)$$

où, $r(b, a) = \sum_{s \in S} r(s, a) b(s)$.

La résolution d'un POMDP devient rapidement très complexe avec le nombre d'états du problème (P-Space), car la solution optimale est recherchée sur l'ensemble d'états de croyance qui est un ensemble continu, en tant que distribution de probabilité b, sur les états réels du système auxquels l'agent n'a pas directement accès. Quand la fonction de valeur est itérée sur un horizon infini, le problème devient indécidable [10, 6] ; on trouve donc des solveurs POMDP qui approchent la solution optimale, par exemple en itérant uniquement sur l'ensemble d'états de croyances *atteignables* depuis un état de croyance initial connu. On peut citer entre autres les algorithmes : Point Based Value Iteration - PBVI [8], PERSEUS [12] et sa version symbolique, Symbolic-PERSEUS [9]. Ce dernier modélise l'espace d'états sous forme de variables d'état et la dynamique probabiliste du système sous forme de Réseaux Bayésiens Dynamiques (DBN) [3]. Il utilise des heuristiques pour limiter le nombre d'états explorés durant l'optimisation du problème. Nous avons choisi d'adapter ce planificateur à nos deux nouveaux critères mixtes, présentés dans la suite.

2.2 Perception Active

Les robots mobiles doivent agir en dépit des incertitudes sur leur environnement et leurs actions. Ils perçoivent leur environnement grâce à des capteurs divers, dont ils fusionnent l'information afin de reconstituer au mieux l'état réel de l'environnement. La perception active [1, 4, 7] peut être définie par l'optimisation de séquences de décisions de l'agent autonome pour observer et fusionner au mieux l'information des différents capteurs, en tenant compte des effets de ses actions sur ses capacités de perception de l'environnement [11].

Choisir les actions de perception requiert un compromis entre actions de court et de long terme. Le robot doit prendre les décisions qui vont maximiser ses chances d'atteindre les buts de sa mission, et les décisions qui vont lui permettre de prendre plus d'informations en relation avec son environnement. Un critère de performance, appelé *fonction de valeur*, est nécessaire pour quantifier chaque séquence d'actions $(a_1, a_2, ..., a_n)$ et le gain attribué à l'acquisition d'informations et celui associé à l'accomplissement des buts de la mission :

$$J = \min_{a_1, \ldots, a_n} \left\{ \sum_j \alpha_j U_j + \sum_l \beta_l C_l \right\} \quad (4)$$

Il peut être composé par la somme de deux récompenses :

- Les termes en j, représentent l'espérance de l'incertitude par rapport à l'état réel du système ; ou encore, l'incertitude liée à la précision nécessaire pour atteindre le but ;
- Les termes en l, qui représentent l'espérance d'autres coûts associés aux déplacements du robot : énergie, temps, distances des obstacles, distance au but.

U_j et C_l, sont tous deux fonctions de la politique $a_1, ..., a_n$. Les pondérations α_j et β_l donnent un poids différent pour les deux termes, et sont des paramètres réglés arbitrairement par le concepteur.

Entropie de croyance. Dans le cadre probabiliste, l'estimation de l'incertitude est basée sur une distribution de probabilité sur les états atteignables du système. Pour mesurer l'incertitude de la distribution de probabilité, il est courant d'utiliser l'entropie de Shannon, qui mesure la quantité d'information dans une distribution de probabilité :

$$H(b_t) = \sum_{s \in S} b_t(s) . \log_m(b_t(s)) \quad (5)$$

où, S est un ensemble de n états discrets, $b_t(s)$ la probabilité de l'état $s \in S$ à l'instant t, et m la base du loga-

rithme. L'entropie est minimale pour les n distributions de probabilité centrées chacune sur un état, c'est-à-dire quand l'agent autonome croît connaître parfaitement l'état réel du système. Plus l'entropie est grande, plus la distribution de probabilité est uniforme sur les états possibles de l'environnement et contient donc peu d'information.

Critères de performance liés à l'incertitude. Dans [1], la localisation active est présentée avec l'objectif d'estimer la position du robot à partir de données extraites de capteurs. L'idée clef de cette approche est que l'efficacité de la localisation est améliorée par les commandes actives de la direction des déplacements du robot et de ces capteurs. Le principe est de contrôler les actionneurs du robot de sorte à minimiser l'espérance de l'incertitude de la distribution de probabilité sur les positions possibles $Bel(l)$ *dans le futur immédiat*, modélisée par l'espérance de l'entropie de l'état de croyance $E_a(H)$, et le coût du déplacement en jeu $v(a)$, *avec un horizon de raisonnement de 1* :

$$a^* = \arg\min_a (E_a(H) + \alpha v(a)), \text{ avec } \alpha \geqslant 0 \text{ et} \quad (6)$$

$$H = -\int Bel(l) \log Bel(l) dl \quad (7)$$

Dans [4], un agent autonome disposant d'une caméra contrôlable doit classifier des données dans un environnement inconnu, en choisissant au mieux des points de prise de vue. Il doit éviter des prises de vue ambiguës, ou exclure certaines hypothèses d'identification. La modélisation du problème attribue une récompense plus importante au choix de point de vue qui augmente la quantité d'information acquise afin de diminuer l'incertitude. La quantité d'information acquise est exprimée par l'entropie de croyance définie précédemment. A l'instant t, la décision, c'est-à-dire le choix du point de vue, aura pour but de maximiser l'espérance cumulée et pondérée des récompenses futures. La récompense, ici, ne dépend pas des coûts liés aux mouvements de la caméra, mais seulement de la quantité d'information acquise :

$$\pi^*(s) = \arg\max_\pi E\left[R_t \mid s_t = s, \pi\right], \text{ avec} \quad (8)$$

$$R_t = -\sum_{n=0}^{\infty} \gamma^n H^\pi(s_{t+n+1}) \quad (9)$$

où $H^\pi(s_t)$ est l'entropie qui mesure l'incertitude de l'état à l'instant t. Deux autres approches sont présentées dans [2] : la première calcule une politique dont l'exécution est contrôlée par la valeur immédiate de $H(b)$, où $H(b)$ est la valeur de l'entropie de l'état de croyance. La deuxième pondère la valeur espérée, soit $V'(b')$ par $H(b')$ sur un horizon immédiat. Aucune des ces approches raisonnent à long terme.

Ainsi, à notre connaissance, il n'existe pas de travaux sur la décision séquentielle (horizon supérieur à 1) en environnement partiellement observable mêlant à la fois des objectifs de perception (gain d'information) et de mission (récompenses de buts) : soit l'un de ces critères a été étudié séparément en horizon supérieur à 1, soit ces deux critères ont été optimisés simultanément mais sur un horizon de 1. Or, il nous semble pertinent d'étudier l'optimisation de ces deux critères simultanément sur un horizon supérieur à 1 pour deux raisons essentielles :

– le gain d'information et les récompenses devraient être *explicitement* optimisés l'un et l'autre en tenant compte des conséquences à long terme des actions choisies ;
– l'optimisation explicite du gain d'information devrait réduire les erreurs de croyance de l'agent et donc augmenter les récompenses cumulées lors de l'exécution de la politique.

Dans la suite, nous présentons deux critères mixtes, multiplicatif puis additif, et les modifications que nous avons apportées au planificateur Symbolic-PERSEUS [9]. Nous montrerons des résultats expérimentaux qui confirment les intuitions soulevées dans les deux items précédents.

3 Critères d'optimisation mixtes

Dans cette section, nous reprenons la mesure du gain d'information à l'aide de l'entropie de Shannon (cf. équation 5, utilisée en perception active [1, 4]. Nous utilisons le logarithme en base 10.

En maximisant les récompenses cumulées tout en minimisant explicitement l'entropie de croyance de l'agent, nous souhaitons éliminer l'incertitude sur son état de croyance plus rapidement qu'en optimisant uniquement les récompenses à long terme des buts de la mission[2].

3.1 Critère d'optimisation multiplicatif

Une de nos approches traite un critère d'optimisation multiplicatif : nous modélisons l'influence de l'espérance de la somme pondérée des récompenses des actions, pondérés 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.

$$J^\pi(b) = E_\pi\left[\sum_{t=0}^{\infty} \gamma^t \frac{r(b_t, \pi(b_t))}{|H(b_t)|} \mid b_0 = b\right] \quad (10)$$

Théorème : équation de Bellman du critère multiplicatif. *La fonction de valeur optimale du critère multiplicatif est la limite de la suite vectorielle définie par :*

$$J_{n+1}(b) = \max_{a \in A}\left\{\frac{r(b,a)}{|H(b)|} + \gamma \sum_{o \in \Omega} p(o|b,a) J_n(b_a^o(s'))\right\} \quad (11)$$

Démonstration. 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 l'inverse des entropies. □

Cette nouvelle équation de Bellman nous permet de calculer par programmation dynamique une politique qui pon-

[2] Notons que la maximisation des récompenses cumulées n'implique pas nécessairement la minimisation de l'incertitude sur l'état de croyance.

dère la récompense immédiate par l'inverse de la valeur absolue de l'entropie de l'état de croyance.

Heuristique. Symbolic-PERSEUS utilise une heuristique numérique pour déterminer l'ensemble des états de croyance initiaux pertinents. Il initialise la recherche des états de croyance atteignables par $V^{\pi}_{degrad} = \max_{s,a} r(s,a)$, calculé 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 les nouveau critère J^{π}, c'est-à-dire un J_0.

Théorème : heuristique du critère multiplicatif. *Une heuristique admissible pour le critère multiplicatif est :*

$$J_0 = \frac{V^{\pi}_{degrad}}{|\log_{10}(n)|(1-\gamma)} \tag{12}$$

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

$$H(b)_{min} = n\frac{1}{n}\log_{10}\left(\frac{1}{n}\right) = -\log_{10}(n) \tag{13}$$

$$\text{Ainsi : } J^{\pi} \geq \frac{V^{\pi}_{degrad}}{|-\log_{10}(n)|(1-\gamma)} = J_0 \tag{14}$$

□

3.2 Critère d'optimisation additif

Le deuxième critère que nous proposons est additif : nous modélisons l'espérance de la somme pondérée des récompenses attribuées aux actions choisies ajoutée à l'espérance de la somme pondérée des gains d'information à long terme, représentés par les entropies de croyance stochastiques successifs. Ces deux valeurs sont elles-mêmes pondérées par deux constantes β et ρ :

$$\begin{aligned} J^{\pi}(b) &= \beta V^{\pi}(b) + \rho H^{\pi}(b), \text{ avec} \\ V^{\pi}(b) &= E_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(b_t, \pi(b_t)) \mid b_0 = b\right] \\ H^{\pi}(b) &= E_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t H(b_t) \mid b_0 = b\right] \end{aligned} \tag{15}$$

Théorème : équation de Bellman du critère additif. *La fonction de valeur optimale du critère additif est la limite de la suite vectorielle définie par :*

$$J_{n+1}(b) = \max_{a \in A}\{\beta.r(b,a) + \rho.H(b) + \gamma \sum_{o \in \Omega} p(o|b,a) J_n(b^o_a(s'))\} \tag{16}$$

Démonstration. L'équation 16 peut se récrire :

$$J^{\pi}(b) = E_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t \left(\beta r(b_t, \pi(b_t)) + \rho H(b_t)\right) \mid b_0 = b\right] \tag{17}$$

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. □

Heuristique. Comme dans le cas multiplicatif, 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 du critère additif. *Une heuristique admissible pour le critère additif est donnée par :*

$$J_0 = \frac{\beta V^{\pi}_{degrad} - \rho \log_{10}(n)}{1-\gamma} \tag{18}$$

Remarque. Les critères présentés ici ne sont pas linéaires par morceaux, mais des algorithmes tels que PBVI [8], PERSEUS [12] et Symbolic-PERSEUS [9], qui approximent le critère par génération stochastique d'états de croyances locaux, peuvent approximer ces critères non-linéaires, sachant que toute fonction continue est approchable par une fonction linéaire par morceaux.

4 Résultats expérimentaux

4.1 Scénario d'étude

Le modèle étudié traite d'un hélicoptère autonome qui cherchera à identifier et pister deux cibles mobiles. Ces cibles sont de natures différentes, l'une est *amie* et l'autre non. Le but de l'hélicoptère autonome est de se poser sur la cible *amie*, 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 se poser sur la bonne.

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. On souligne que, pour les simulations étudiées dans ce travail, nous avons donné un état de

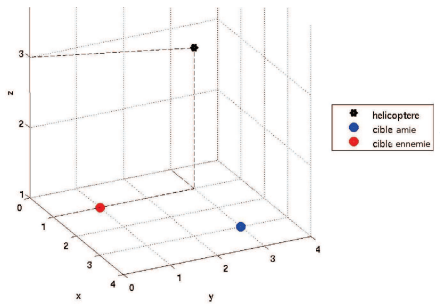

FIG. 1 – Position de départ de l'hélicoptère autonome et des cibles 1 (amie) et 2 (ennemie).

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 *amie* et la cible 2, *ennemie*. L'espace de déplacement de

Cibles	Amie	Ennemie
cible 1	0.2 ($\times$)	0.8
cible 2	0.8	0.2 ($\times$)

TAB. 1 – État de croyance initial de l'agent sur les cibles ; le signe ($\times$) indique la nature réelle des cibles.

l'hélicoptère est modélisé par une grille $3 \times 3 \times 3$, et celui des cibles $3 \times 3 \times 1$, 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 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 à l'hélicoptère, seulement s'il est au-dessus d'une cible et à une altitude de 1 par rapport à la cible ($z = 2$). Une fois que l'hélicoptère a atterri, sur la cible *amie*, ou sur la cible *ennemie*, 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 *amie*.
Le modèle d'observation de la nature des cibles dépend de la distance géométrique entre l'hélicoptère et les cibles :

$$\begin{aligned} p(o' = \text{amie} \mid s' = \text{amie}) &= \\ p(o' = \text{ennemie} \mid s' = \text{ennemie}) &= \frac{1}{2}\left(e^{\frac{-d}{D}} + 1\right) \quad (19) \\ p(o' = \text{amie} \mid s' = \text{ennemie}) &= \\ p(o' = \text{ennemie} \mid s' = \text{amie}) &= \frac{1}{2}\left(1 - e^{\frac{-d}{D}}\right) \quad (20) \end{aligned}$$

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.

4.2 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* (et pessimiste) 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'optimalité 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 21.

$$V^{\pi}(s_t) = E^{\pi}_{100\, simulations}\left[\sum_{k=t}^{50} \gamma^k r^{\pi}(s_k) \mid s_t\right] \quad (21)$$

Nous comparerons 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.
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 22.

$$H^{\pi}(b_t) = E^{\pi}_{100\, simulations}\left[\sum_{k=t}^{50} \gamma^k H^{\pi}(b_k) \mid b_t\right] \quad (22)$$

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

4.3 Résultat des simulations

Critère Multiplicatif. La figure 2 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 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 3 : 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.

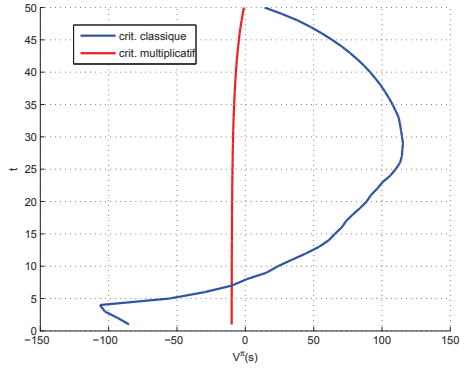

FIG. 2 – Moyenne statistique de la fonction de valeur de l'état courant $V^{\pi}(s_t)$ (l'état change le long de la courbe).

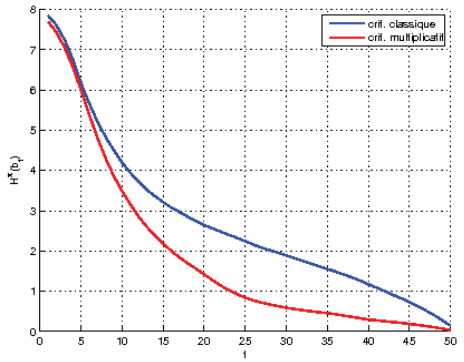

FIG. 3 – Moyenne statistique de la somme de l'entropie pondérée de l'état de croyance courant $H^{\pi}(b_t)$.

Critère Additif. Des politiques ont été calculées pour différentes valeurs du couple de coefficients (β, ρ) : $(1,0;0,0)$, $(0,5;0,5)$ et $(0,0;1,0)$. Le cas $(1,0;0,0)$ 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 d'information, le troisième optimise le gain d'information.

Dans la figure 4, la moyenne des fonctions de valeur $V^{\pi}(s)$ (équation 21), pour les trois cas sont montrés. Dans le premier cas, $\beta = 1.0$ et $\rho = 0.0$, la fonction de valeur statistique $V^{\pi}(s)$ part d'une valeur négative, ce qui s'explique grâce à 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 *ennemie* au bout de 3 ou 4 pas de simulation, car il croit à ce moment que cette cible est la cible *amie*. Par contre, pour les simulations où l'hélicoptère a pu acquérir plus d'information de son environnement, il est vérifié qu'en moyenne, l'hélicoptère autonome inverse son état de croyance (non montré dans cet article pour raison de place) 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, $\beta = 0.5$ et $\rho = 0.5$, 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 démontrée, puisque l'agent cherche maintenant de façon explicite à acquérir plus d'information de son environnement, ce qui lui permet d'inverser plus tôt sont 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 4 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, $\beta = 0.0$ et $\rho = 1.0$, le critère optimise le gain d'information uniquement. La figure 4 montre que la moyenne de $V^{\pi}(s)$ reste proche de zéro : l'hélicoptère ne cherche pas atterrir, car le gain d'atterrissage modélisé dans les récompenses n'est pas pris en compte.

La figure 5 compare le critère subjectif $H^{\pi}(b)$, équation 22, 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 5 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, intégrées 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.

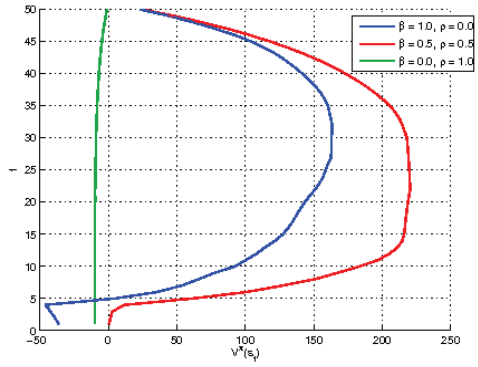

FIG. 4 – Moyenne statistique de la fonction de valeur de l'état courant $V^{\pi}(s_t)$ (l'état change le long de la courbe).

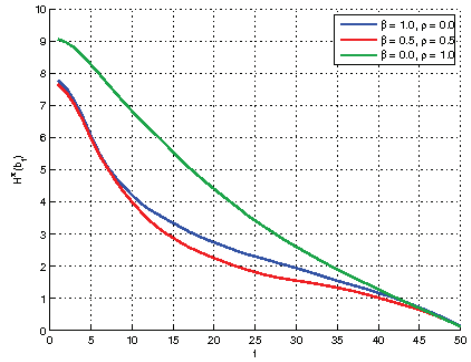

FIG. 5 – Moyenne statistique de la somme de l'entropie pondérée de l'état de croyance $H^{\pi}(b_t)$.

5 Conclusion

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 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'information, 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'information explicite. En quelques sortes, 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 souhaitons étudier plus finement l'influence des coefficients de pondération du gain d'information et des récompenses dans le critère additif, afin de trouver le couple de pondérations qui maximise la somme de récompenses objectivement reçue lors de l'exécution de la politique optimisée.

Références

[1] W. BURGARD, Dieter FOX et Sebastian THRUN : Active mobile robot localization. *In Proceedings of IJCAI-97*. Morgan Kaufmann, 1997.

[2] A.R. CASSANDRA, L.P. KAELBLING et J.A. KURIEN : Acting under uncertainty : Discrete Bayesian models for mobile-robot navigation. *In In Proceedings of IEEE/RSJ*, 1996.

[3] T. DEAN et K. KANAZAWA : A model for reasoning about persistence and causation. *Computational Intelligence*, 5(3):142–150, 1990.

[4] F. DEINZER, J. DENZLER et H. NIEMANN : Viewpoint selection-planning optimal sequences of views for object recognition. *Lecture notes in computer science*, pages 65–73, 2003.

[5] L. P. KAELBLING, M. L. LITTMAN et A. R. CASSANDRA : Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101:99–134, 1998.

[6] O. MADANI, S. HANKS et A. CONDON : On the undecidability of probabilistic planning and related stochastic optimization problems. *Artificial Intelligence*, 147(1):5–34, 2003.

[7] L. MIHAYLOVA, T. LEFEBVRE, H. BRUYNINCKX, K. GADEYNE et J. De SCHUTTER : Active sensing for robotics – a survey. *In 5th Intl Conf. On Numerical Methods and Applications*, pages 316–324, 2002.

[8] J. PINEAU, G. GORDON et S. THRUN : Point-based value iteration : An anytime algorithm for POMDPs. *In International Joint Conference on Artificial Intelligence*, volume 18, pages 1025–1032. LAWRENCE ERLBAUM ASSOCIATES LTD, 2003.

[9] P. POUPART : *Exploiting structure to efficiently solve large scale partially observable Markov decision processes*. Thèse de doctorat, University of Toronto, 2005.

[10] O. SIGAUD et O. BUFFET : *Processus Décisionnels de Markov en intelligence artificielle*, volume 1. Lavoisier and Hermes Sciences, 2008.

[11] M.T.J. SPAAN : Cooperative Active Perception using POMDPs. *Association for the Advancement of Artificial Intelligence - AAAI*, 2008.

[12] M.T.J. SPAAN et N. VLASSIS : A point-based POMDP algorithm for robot planning. *In IEEE International Conference on Robotics and Automation*, volume 3, pages 2399–2404. IEEE ; 1999, 2004.