Une extension des POMDP avec des récompenses dépendant de l'état de croyance

Mauricio Araya-López, Olivier Buffet, Vincent Thomas, François Charpillet

Nancy Université / INRIA
LORIA – Campus Scientifique – BP 239
54506 Vandoeuvre-lès-Nancy Cedex – France
prénom.nom@loria.fr

[Article publié en anglais dans Advances in Neural Processing Systems 24*.]*

Résumé : Les processus de décision markoviens partiellement observables (POMDP) modélisent des problèmes de prise de décision séquentielle dans l'incertain et sous observation partielle. Toutefois, certains problèmes ne peuvent être modélisés avec des fonctions de récompense dépendant de l'état, comme des problèmes dont l'objectif requiert explicitement de réduire l'incertitude sur l'état. Dans ce but, nous introduisons les ρPOMDP, une extensions des POMDP dans laquelle la fonction de récompense ρ dépend de l'état de croyance. Nous montrons que, sous l'hypothèse courante que ρ est convexe, la fonction de valeur est aussi convexe, ce qui permet (1) d'approcher ρ arbitrairement bien avec une fonction convexe et linéaire par morceaux (PWLC), et (2) d'employer des algorithmes de résolutions de l'état de l'art, exacts ou approchés, avec des modifications minimales.

Mots-clés : processus de décision markoviens partiellement observables, fonction de récompense, perception active, approximation convexe et linéaire par morceaux

1. Introduction

Les problèmes de prise de décision séquentielle dans l'incertain et sous observations partielles sont typiquement modélisés à l'aide de processus de décision markoviens partiellement observables (POMDP) (Smallwood & Sondik, 1973), dans lesquels l'objectif est de décider comment agir de manière à ce que la séquence d'états visités optimise un critère de performance donné. Toutefois, ce formalisme n'est pas assez expressif pour modéliser des problèmes avec n'importe quel type de fonction objectif.

Considérons des problèmes d'*active sensing* (perception active), où l'objectif est d'acquérir de la connaissance à propos de certaines variables d'état. Le diagnostic médical par exemple consiste à poser les bonnes questions et effectuer les examens appropriés de manière à diagnostiquer un patient à faible coût avec une haute certitude. Ceci peut être formalisé par un POMDP en récompensant – en cas de succès – une action finale consistant à exprimer la "meilleure hypothèse" du diagostic. En fait, de nombreux travaux formalisent l'*active sensing* avec des POMDP (Thrun, 2000; Mihaylova *et al.*, 2002; Ji & Carin, 2007).

Une difficulté est que, dans certains problèmes, l'objectif a besoin d'être exprimé directement en termes d'incertitude/information sur l'état, par exemple minimiser l'entropie sur une certaine variable d'état. Dans de tels cas, les POMDP ne sont pas appropriés parce que la fonction de récompense ne dépend que de l'état et de l'action et pas des connaissances de l'agent. Au lieu de cela, nous avons besoin d'un modèle dans lequel la récompense immédiate dépend de l'*état de croyance* courant. Le formalisme des *belief MDP* fournit l'expressivité requise pour ces problèmes. Il y a cependant peu de recherches sur des algorithmes spécifiques pour les résoudre, de sorte qu'on les force usuellement à rentrer dans le cadre des POMDP, ce qui implique de changer la définition originale du problème. On peut argumenter qu'acquérir de l'information est toujours un moyen et non une fin, et donc qu'un problème de prise de décision séquentielle avec observabilité partielle "bien défini" doit toujours être modélisé comme un POMDP normal. Toutefois, dans un certain nombre de cas le concepteur du problème a décidé de séparer les tâches de recherche et d'exploitation de l'information. Citons à ce titre deux exemples : (i) la surveillance (Hero *et al.*, 2007) et (ii) l'exploration (Thrun, 2000) d'une zone donnée, dans les deux cas sans savoir ce qui est attendu de ces tâches – et donc sans savoir comment réagir à des découvertes.

Après quelques rappels sur les POMDP en section 2., la section 3. introduit les ρPOMDP– une extension des POMDP dans laquelle la récompense est une fonction (typiquement convexe) de l'état de croyance – et prouve que la convexité de la fonction de valeur est préservée. Ensuite nous montrons comment des algorithmes de résolution classiques peuvent être adaptés selon que la fonction de récompense est linéaire par morceaux (sec. 3.3.) ou non (sec. 4.).

2. MDP Partiellement Observables

Le problème général auquel les POMDP s'attaquent est, pour un agent, de trouver une *politique* de décision π choisissant, à chaque pas de temps, la meilleure action en fonction de ses observations et actions passées de manière à maximiser son gain futur (lequel peut être mesuré par exemple à travers la récompense totale accumulée ou la récompense moyenne par pas de temps). En comparaison avec la planification déterministe classique, l'agent doit faire face à une difficulté supplémentaire : non seulement le système à contrôler a une dynamique incertaine, mais son état courant est en plus imparfaitement connu.

2.1. Description d'un POMDP

Formellement, les POMDP sont définis par un tuple $\langle \mathcal{S}, \mathcal{A}, \Omega, T, O, r, b_0 \rangle$ où, à chaque pas de temps, le système étant dans un état $s \in \mathcal{S}$ (l'*espace d'états*), l'agent effectue une action $a \in \mathcal{A}$ (l'*espace d'actions*) qui résulte en (1) une transition vers l'état s' selon la *fonction de transition* $T(s, a, s') = Pr(s'|s, a)$, (2) une observation $o \in \Omega$ (l'*espace d'observations*) selon la *fonction d'observation* $O(s', a, o) = Pr(o|s', a)$, et (3) une *récompense* scalaire $r(s, a)$. b_0 est la distribution de probabilité initiale sur les états. Sauf indication contraire, les ensembles d'états, d'actions et d'observations sont finis (Cassandra, 1998).

L'agent peut typiquement raisonner sur l'état du système en calculant un *état de croyance* $b \in \Delta = \Pi(\mathcal{S})$ (l'ensemble des distributions de probabilité sur $\mathcal{S}$), [1] en utilisant la formule de mise à jour suivante (reposant sur la loi de Bayes) quand l'action a est exécutée et suivie de l'observation o :

$$b^{a,o}(s') = \frac{O(s', a, o)}{Pr(o|a, b)} \sum_{s \in \mathcal{S}} T(s, a, s') b(s),$$

où $Pr(o|a, b) = \sum_{s,s'' \in \mathcal{S}} O(s'', a, o) T(s, a, s'') b(s)$.

A l'aide des états de croyance, un POMDP peut être ré-écrit comme un MDP sur l'espace des croyances, appelé *belief MDP*, $\langle \Delta, \mathcal{A}, \tau, \rho \rangle$, où les nouvelles fonctions de transition τ et de récompense ρ sont définies respectivement sur $\Delta \times \mathcal{A} \times \Delta$ et $\Delta \times \mathcal{A}$. Avec cette reformulation, plusieurs

1. $\Pi(\mathcal{S})$ forme un simplexe parce que $\|b\|_1 = 1$, c'est pourquoi nous utilisons Δ comme l'ensemble de tous les b possibles.

résultats théoriques sur les MDP peuvent être étendus, tels que l'existence d'une politique déterministe qui soit optimale. Une difficulté est que, même si un POMDP a un nombre fini d'états, le *belief MDP* correspondant est défini sur un espace de croyance continu – et donc infini.

Dans ce MDP continu, l'objectif est de maximiser les récompenses cumulées en cherchant une politique prenant comme entrée l'état de croyance courant. Plus formellement, nous cherchons une politique vérifiant $\pi^* = \text{argmax}_{\pi \in \mathcal{A}^\Delta} J^\pi(b_0)$ où $J^\pi(b_0) = E\left[\sum_{t=0}^{\infty} \gamma^t \rho_t | b_0, \pi\right]$, ρ_t étant la récompense immédiate espérée obtenue au temps t, et γ un facteur d'actualisation. Le principe d'optimalité de Bellman (Bellman, 1954) permet de calculer la fonction J^{π^*} récursivement à travers la *fonction de valeur*

$$\begin{aligned} V_n(b) &= \max_{a \in \mathcal{A}} \left[\rho(b,a) + \gamma \int_{b' \in \Delta} \tau(b,a,b') V_{n-1}(b') db' \right] \\ &= \max_{a \in \mathcal{A}} \left[\rho(b,a) + \gamma \sum_o Pr(o|a,b) V_{n-1}(b^{a,o}) \right], \end{aligned} \tag{1}$$

où, pour tout $b \in \Delta$, $V_0(b) = 0$, et $J^{\pi^*}(b) = V_{n=H}(b)$ (où H est l'horizon – éventuellement infini – du problème).

Le cadre des POMDP emploie une fonction de récompense $r(s,a)$ reposant sur l'état et l'action. Le belief MDP, pour sa part, emploie une fonction de récompense $\rho(b,a)$ reposant sur l'état de croyance. Cette fonction de récompense dépendant de la croyance est dérivée comme une espérance sur la récompense du POMDP :

$$\rho(b,a) = \sum_s b(s) r(s,a). \tag{2}$$

Une conséquence importante de l'équation 2 est que le calcul récursif décrit en équation 1 a la propriété de générer des fonctions de valeur convexes et linéaires par morceaux (PWLC) pour chaque horizon (Smallwood & Sondik, 1973). Chaque fonction est ainsi déterminée par un ensemble d'hyperplans (chacun représenté par un vecteur), la valeur à un état de croyance donné étant celle de l'hyperplan le plus haut. Par exemple, si Γ_n est l'ensemble des vecteurs représentant la fonction de valeur à l'horizon n, alors $V_n(b) = \max_{\alpha \in \Gamma_n} \sum_s b(s)\alpha(s)$.

2.2. Résolution de POMDP avec des mises à jours exactes

En utilisant la propriété PWLC, on peut effectuer la mise à jour de Bellman en employant la factorisation suivante de l'équation 1 :

$$V_n(b) = \max_{a \in \mathcal{A}} \sum_o \sum_s b(s) \left[\frac{r(s,a)}{|\Omega|} + \gamma \sum_{s'} T(s,a,s') O(s',a,o) \chi_{n-1}(b^{a,o}, s') \right], \tag{3}$$

avec [2] $\chi_n(b) = \underset{\alpha \in \Gamma_n}{\operatorname{argmax}}\, b \cdot \alpha$. Si on considère le terme entre crochets dans l'équation 3, cela génère $|\Omega| \times |\mathcal{A}|$ Γ-sets, chacun de taille $|\Gamma_{n-1}|$. Ces ensembles sont définis comme

$$\overline{\Gamma_n}^{a,o} = \left\{ \frac{r^a}{|\Omega|} + P^{a,o} \cdot \alpha_{n-1} \,\middle|\, \alpha_{n-1} \in \Gamma_{n-1} \right\}, \tag{4}$$

où $P^{a,o}(s,s') = T(s,a,s')O(s',a,o)$ et $r^a(s) = r(s,a)$. Ainsi, pour obtenir une représentation exacte de la fonction de valeur, on peut calculer ($\bigoplus$ étant la somme croisée de deux ensembles) :

$$\overline{\Gamma_n} = \bigcup_a \bigoplus_o \overline{\Gamma_n}^{a,o}.$$

Cependant, ces ensembles $\overline{\Gamma_n}^{a,o}$ – ainsi que le $\overline{\Gamma_n}$ final – sont *non parcimonieux* : certains α-vecteurs peuvent être inutiles parce que les hyperplans correspondants se situent sous la fonction de valeur. Des phases d'élagage sont alors requises pour retirer les vecteurs dominés. Plusieurs algorithmes reposent sur des techniques d'élagage, tels que *Batch Enumeration* (Monahan, 1982) ou des algorithmes plus efficaces comme *Witness* ou *Incremental Pruning* (Cassandra, 1998).

2.3. Résolution de POMDP avec des mises à jour approchées

Les procédés de mise à jour de la fonction de valeur présentés ci-dessus sont exacts et fournissent des fonctions de valeur qui peuvent être utilisées quel que soit l'état de croyance initial b_0. Diverses solutions approchées de POMDP ont été proposées pour réduire la complexité de ces calculs, employant par exemple des estimations heuristiques de la fonction de valeur, ou

2. La fonction χ retourne un vecteur, de sorte que $\chi_n(b,s) = (\chi_n(b))(s)$.

appliquant la mise à jour de la valeur seulement sur des états de croyance choisis (Hauskrecht, 2000). Nous nous focalisons ici sur ces dernières approximations dites *à base de points* (*point-based*) (PB), lesquelles ont largement contribué aux récents progrès dans la résolution de POMDP, et dont la littérature associée va des premiers travaux par Lovejoy (Lovejoy, 1991) à SARSOP par Kurniawati *et al.* (Kurniawati *et al.*, 2008) en passant par PBVI par Pineau *et al.* (Pineau *et al.*, 2006), Perseus par Spaan et Vlassis (Spaan & Vlassis, 2005), et HSVI2 par Smith et Simmons (Smith & Simmons, 2005).

A chaque itération n jusqu'à convergence, un algorithme PB typique :

1. sélectionne un nouvel ensemble d'états de croyance B_n en fonction de B_{n-1} et de l'approximation courante V_{n-1} ;
2. effectue une mise à jour de Bellman à chaque état de croyance $b \in B_n$, produisant un α-vecteur par point ; et
3. élague les points dont les hyperplans associés sont dominés ou considérés comme négligeables.

Les multiples algorithmes PB diffèrent principalement par la façon de choisir les états de croyance et la façon d'effectuer la mise à jour. Les méthodes existantes de sélection des états de croyance exploitent des idées telles que l'utilisation d'une discrétisation régulière ou un échantillonnage aléatoire du simplexe, choisissant des points atteignables (en simulant des séquences d'actions commençant de b_0), et ajoutant les points qui réduisent l'erreur d'approximation, ou en cherchant en particulier des régions pertinentes pour la politique optimale (Kaplow, 2010).

3. Une extension des POMDP pour l'active sensing

3.1. Introduction des ρPOMDP

Tous les problèmes avec observabilité partielle rencontrent la difficulté d'obtenir plus d'information pour atteindre un but. Usuellement, ce problème est explicitement traité dans le processus de résolution, où l'acquisition d'information n'est qu'un moyen pour optimiser une récompense espérée dépendant de l'état du système. Certains problèmes d'active sensing peuvent être modélisés de cette manière (par exemple la classification active), mais pas tous. Ainsi, on rencontre aussi des problèmes dans lesquels le critère de performance inclut une mesure explicite de la connaissance de l'agent à propos du système, ce qui repose sur les croyances et non sur les états. La surveillance

par exemple est une tâche sans fin qu'il ne semble pas possible de modéliser avec des récompenses dépendant de l'état. En effet, si on considère le simple problème de déterminer la position d'un objet caché, celui-ci peut être résolu sans même avoir vu cet objet (par exemple si toutes les localisations possibles sauf une ont été visitées). Toutefois, la fonction de récompense d'un POMDP ne peut modéliser ceci puisqu'elle repose sur l'état et l'action courants. Une solution serait d'inclure tout l'historique dans l'état, avec pour conséquence une explosion combinatoire. Nous préférons considérer une nouvelle manière de définir les récompenses reposant sur la connaissance acquise représentée par les états de croyance. La suite de cet article étudie la possibilité d'utiliser les belief MDP hors de la définition spécifique de $\rho(b, a)$ de l'équation 2, et discute de la manière de résoudre ce type spécifique de problèmes d'active sensing.

Comme l'équation 2 n'est plus valide, le lien direct avec les POMDP est rompu. On peut cependant toujours utiliser tous les autres composants des POMDP tels que les états, les observations, etc. L'approche que nous proposons consiste à généraliser le cadre des POMDP aux ρPOMDP, dans lesquels la récompense n'est pas définie comme une fonction $r(s, a)$, mais directement comme une fonction $\rho(b, a)$. La nature de la fonction $\rho(b, a)$ dépend du problème, mais elle est usuellement liée à une mesure d'incertitude ou d'erreur (Mihaylova *et al.*, 2002; Thrun, 2000; Ji & Carin, 2007). La plupart des méthodes communes sont celles reposant sur la théorie de l'information de Shannon, en particulier l'entropie de Shannon ou la distance de Kullback-Leibler distance (Cover & Thomas, 1991). De manière à présenter ces fonctions comme des récompenses, elles doivent mesurer l'information plutôt que l'incertitude, de sorte qu'on emploie la fonction d'entropie négative $\rho_{ent}(b) = \log_2(|S|) + \sum_{s \in S} b(s) \log_2(b(s))$ – qui est maximale dans les coins du simplexe et minimale en son centre – plutôt que l'entropie originale de Shannon. En outre, d'autres fonctions plus simples reposant sur la même idée peuvent être utilisées, telles que la distance au centre du simplexe (DSC pour l'anglais *distance from the simplex center*), $\rho_{dsc}(b) = \|b - c\|_m$, où c est le centre du simplexe et m un entier positif qui dénote l'ordre de l'espace métrique. On notera que $\rho(b, a)$ n'est pas restreint à être seulement une mesure d'incertitude, mais peut aussi être une combinaison de récompenses espérées dépendant de l'état et de l'action – comme dans l'équation 2 – et d'une mesure d'incertitude ou d'erreur. Par exemple, Mihaylova *et al.* (Mihaylova *et al.*, 2002) définissent le problème de l'active sensing comme l'optimisation de la somme pondérée de mesures d'incertitudes et de coûts, le premier dépendant

de la croyance et le second de l'état du système.

Dans la suite de cet article, nous montrons comment appliquer aux ρPOMDP les algorithmes classiques pour POMDP. Dans ce but, nous discutons de la convexité de la fonction de valeur, laquelle permet d'étendre ces algorithmes à l'aide d'approximations PWLC.

3.2. Propriété de convexité

Une propriété importante utilisée pour résoudre les POMDP normaux est que la fonction de valeur définie sur les croyances est convexe. Cela vient du fait que $r(s, a)$ se traduit par une fonction linéaire dans l'espace des croyances, et que les opérateurs espérance, somme et maximum préservent cette propriété (Smallwood & Sondik, 1973). Pour les ρPOMDP, cette propriété est aussi vérifiée si la fonction de récompense $\rho(b, a)$ est convexe, comme le dit le théorème 1.

Theorème 1
Si ρ et V_0 sont des fonctions convexes sur Δ, alors la fonction de valeur V_n du belief MDP est convexe sur Δ à tout pas de temps n. [Preuve dans (Araya-López et al.*, 2010, annexe)]*

Ce théorème se base sur le fait que $\rho(b, a)$ est une fonction convexe sur b, ce qui est une propriété naturelle pour les mesures d'incertitude (ou d'information). En effet, comme l'objectif est d'éviter les distributions qui ne donnent pas beaucoup d'information sur l'état dans lequel se trouve le système, on cherche à affecter de plus hautes récompenses aux états de croyance qui correspondent à une plus grande probabilité d'être dans un état particulier. Ainsi, une fonction de récompense supposée réduire l'incertitude doit fournir d'importants gains près des coins du simplexe, et de faibles gains près du centre. Pour cette raison, nous nous focaliserons sur les fonctions de récompense convexes dans la suite de cet article.

La fonction de valeur initiale V_0 peut être n'importe quelle fonction convexe pour les problèmes à horizon infini, mais par définition $V_0 = 0$ pour les problèmes à horizon fini. Sans perte de généralité, nous ne considérons dorénavant que ce dernier cas. En outre, en partant de $V_0 = 0$, il est aussi aisé de prouver par induction que, si ρ est continue (respectivement dérivable), alors V_n est continue (respectivement dérivable *par morceaux*).

3.3. Fonctions de récompense linéaires par morceaux

Cette section se concentre sur le cas où ρ est une fonction PWLC. Elle montre qu'une petite adaptation des mises à jour exactes et approchées pour le cas des POMDP est nécessaire pour calculer la fonction de valeur optimale. Le cas complexe où ρ n'est pas PWLC sera abordé en section 4..

3.3.1. Mises à jour exactes

Désormais, comme la fonction $\rho(b, a)$ est une fonction PWLC, elle peut être représentée par plusieurs Γ-sets, un Γ_ρ^a pour chaque action a. La récompense peut être calculée par :

$$\rho(b, a) = \max_{\alpha \in \Gamma_\rho^a} \left[\sum_s b(s)\alpha(s) \right].$$

L'usage de cette définition mène aux changements suivants dans l'équation 3

$$V_n(b) = \max_{a \in \mathcal{A}} \sum_s b(s) \left[\chi_\rho^a(b, s) + \sum_o \sum_{s'} T(s, a, s')O(s', a, o)\chi_{n-1}(b^{a,o}, s') \right],$$

où $\chi_\rho^a(b, s) = \underset{\alpha \in \Gamma_\rho^a}{\operatorname{argmax}}(b \cdot \alpha)$. Ceci utilise le Γ-set Γ_ρ^a et génère $|\Omega| \times |A|$ Γ-sets :

$$\overline{\Gamma_n}^{a,o} = \{P^{a,o} \cdot \alpha_{n-1} |\, \alpha_{n-1} \in \Gamma_{n-1}\},$$

où $P^{a,o}(s, s') = T(s, a, s')O(s', a, o)$.

Des algorithmes exacts comme Value Iteration ou Incremental Pruning peuvent être appliqués à cette extension des POMDP d'une manière similaire aux POMDP. La différence est que la somme croisée inclut non seulement un vecteur $\alpha^{a,o}$ pour chaque Γ-set d'observation $\overline{\Gamma_n}^{a,o}$, mais aussi un vecteur α_ρ du Γ-set Γ_ρ^a correspondant à la récompense :

$$\overline{\Gamma_n} = \bigcup_a \left[\bigoplus_o \overline{\Gamma_n}^{a,o} \oplus \Gamma_\rho^a \right].$$

Ainsi, la somme croisée génère $|R|$ fois plus de vecteurs que pour un POMDP classique, $|R|$ étant le nombre d'α-vecteurs décrivant la fonction $\rho(b, a)$.[3]

3. Plus précisément, le nombre $|R|$ dépend de l'action considérée.

3.3.2. Mises à jour approchées

Les approximations à base de points peuvent être appliquées de la même manière que le sont PBVI ou SARSOP. La seule différence est à nouveau que la fonction de récompense est l'enveloppe d'un ensemble d'hyperplans. Comme les algorithmes PB sélectionnent l'hyperplan qui maximise la fonction de valeur à chaque état de croyance, la même simplification peut être appliquée à l'ensemble Γ^a_ρ.

4. Généralisation à d'autres fonctions de récompense

Des mesures d'incertitude comme l'entropie négative ou la distance au centre du simplexe (avec $m > 1$ et $m \neq \infty$) ne sont pas des fonctions linéaires par morceaux. En théorie, chaque étape de value iteration peut être calculée analytiquement en utilisant ces fonctions, mais les expressions ne sont pas en forme fermée comme dans le cas linéaire. Ainsi, les expressions analytiques croissent rapidement en complexité et deviennent ingérables après quelques étapes. De plus, les techniques d'élagage ne peuvent être appliquées directement aux hypersurfaces résultantes, et même des mesures du second ordre ne fournissent pas des formes quadratiques standard auxquelles appliquer la programmation quadratique. Toutefois, les fonctions convexes peuvent être efficacement approchées par des fonctions linéaires par morceaux, ce qui rend possible la mise en œuvre des techniques décrites en section 3.3. avec une erreur bornée tant que l'approximation de ρ est bornée.

4.1. Approximation de ρ

Considérons une fonction de récompense $\rho(b)$ continue, convexe, et dérivable par morceaux,[4] et un ensemble $B \subset \Delta$ arbitraire (et fini) de points où le gradient est défini. Une approximation PWLC par en dessous de $\rho(b)$ peut être obtenue en utilisant chaque élément $b' \in B$ comme un point de référence pour construire un hyperplan tangent qui sera toujours un minorant de $\rho(b)$. Concrètement, $\omega_{b'}(b) = \rho(b') + (b - b') \cdot \nabla\rho(b')$ est la fonction linéaire qui représente l'hyperplan tangent. Ainsi, l'approximation de $\rho(b)$ utilisant un ensemble B est définie par $\omega_B(b) = \max_{b'}(\omega_{b'}(b))$.

4. Par commodité – et sans perte de généralité – nous ne considérons que le cas où $\rho(b, a) = \rho(b)$.

A n'importe quel point $b \in \Delta$, l'erreur de l'approximation peut être écrite

$$\epsilon_B(b) = |\rho(b) - \omega_B(b)|. \tag{5}$$

Si on considère le point spécifique b d'erreur maximale ($= argmax_b(\epsilon_B(b))$), on peut alors essayer de borner cette erreur selon la nature de ρ.

Il est bien connu qu'une approximation linéaire par morceaux d'une fonction lipschtzienne est bornée parce que le gradient $\nabla\rho(b')$ utilisé pour construire l'hyperplan $\omega_{b'}(b)$ a une norme bornée (Saigal, 1979). Malheureusement, l'entropie négative n'est pas lipschitzienne ($f(x) = x\log_2(x)$ a une pente infinie quand $x \to 0$), donc ce résultat n'est pas assez générique pour couvrir une large variété de problèmes d'active sensing. Toutefois, sous certaines hypothèses faibles, il est possible d'établir une borne d'erreur.

L'objectif du reste de cette section est de trouver une borne d'erreur en trois étapes. D'abord, nous allons introduire quelques résultats élémentaires concernant le simplexe et la convexité de ρ. De manière informelle, le lemme 2 montrera que, pour chaque b, il est possible de trouver un état de croyance dans B qui soit assez loin de la frontière du simplexe mais à une distance bornée de b. Ce résultat nous permettra de nous éloigner du cas problématique de la frontière du simplexe tout en nous assurant que nous restons assez proche des points utilisés pour la construction de l'approximation. Ensuite, dans un second temps, nous ferons l'hypothèse que la fonction $\rho(b)$ vérifie la condition α-Hölder pour être capable de borner la norme du gradient dans le lemme 3. Enfin, le théorème 4 utilisera les deux lemmes pour borner l'erreur de l'approximation de ρ sous ces hypothèses.

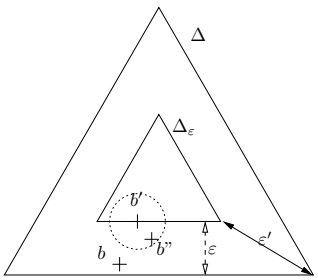

FIGURE 1: Les simplexes Δ et Δ_ε, et les points b, b' et b''.

Pour chaque point $b \in \Delta$, il est possible d'associer un point $b^* = \mathrm{argmax}_{x \in B}\, \omega_x(b)$ correspondant au point de B dont l'hyperplan tangent donne la meilleur approximation de ρ en b. Considérons le point $b \in \Delta$ où l'erreur

d'approximation $\epsilon_B(b)$ est maximale : cette erreur peut être facilement calculée en employant le gradient $\nabla\rho(b^*)$. Malheureusement, certaines dérivées partielles de ρ peuvent diverger vers l'infini sur la frontière du simplexe dans le cas non-lipschitzien, rendant l'erreur difficile à analyser. Ainsi, pour assurer que cette erreur peut être bornée, au lieu de b^* nous allons prendre un $b'' \in B$ sûr (suffisamment loin de la frontière) en utilisant un point intermédiaire b' dans le *simplexe intérieur* Δ_ε, où $\Delta_\varepsilon = \{b \in [\varepsilon, 1]^{\mathcal{N}} | \sum_i b_i = 1\}$ avec $\mathcal{N} = |\mathcal{S}|$.

Ainsi, pour un $b \in \Delta$ donné et $\varepsilon \in (0, \frac{1}{\mathcal{N}}]$, nous définissons le point $b' = \text{argmin}_{x \in \Delta_\varepsilon} \|x - b\|_1$ comme le point le plus près de b dans Δ_ε et $b'' = \text{argmin}_{x \in B} \|x - b'\|_1$ comme le point le plus près de b' dans B (voir figure 1). Ces deux points seront utilisés pour trouver un majorant de la distance $\|b - b''\|_1$ reposant sur la *densité* de B, définie comme $\delta_B = \min_{b \in \Delta} \max_{b' \in B} \|b - b'\|_1$.

Lemme 2
La distance (norme 1) entre le point d'erreur maximale $b \in \Delta$ et le point $b'' \in B$ sélectionné est majorée par $\|b - b''\|_1 \leq 2(\mathcal{N} - 1)\varepsilon + \delta_B$. [Preuve dans (Araya-López et al.*, 2010, annexe)]*

Si on prend $\varepsilon > \delta_B$, on est alors sûr que b'' n'est pas sur la frontière du simplexe Δ, la distance minimale à la frontière étant $\eta = \varepsilon - \delta_B$. Cela va permettre de trouver des bornes pour l'approximation PWLC de fonctions α-Hölder convexes, une famille plus large de fonctions incluant l'entropie négative, des fonctions lipschitziennes convexes et d'autres. La condition α-Hölder est une généralisation de la condition de Lipschitz. Dans notre cadre elle signifie, pour une fonction $f : \mathcal{D} \mapsto \mathbb{R}$ avec $\mathcal{D} \subset \mathbb{R}^n$, qu'elle satisfait

$$\exists \alpha \in (0, 1], \exists K_\alpha > 0, \text{ s.t. } |f(x) - f(y)| \leq K_\alpha \|x - y\|_1^\alpha.$$

Le cas limite, où une fonction α-Hölder convexe a un gradient de norme infinie, se situe toujours sur la frontière du simplexe Δ (du fait de la convexité), et ainsi le point b'' sera protégé de cette situation fâcheuse grâce à η. Plus précisément, une fonction α-hölderienne sur Δ de constante K_α en norme 1 satisfait la condition de Lipschitz sur Δ_η avec la constante $K_\alpha \eta^\alpha$ (voir (Araya-López *et al.*, 2010, annexe)). En outre, la norme du gradient $\|\nabla f(b'')\|_1$ est aussi bornée comme le dit le lemme 3.

Lemme 3
Soit $\eta > 0$ et f une fonction α-hölderienne (de constante K_α), bornée et convexe de Δ dans $\mathbb{R}$, f étant dérivable partout sur Δ^o (l'intérieur de Δ).

Alors, pour tout $b \in \Delta_\eta$, $\|\nabla f(b)\|_1 \leq K_\alpha \eta^{\alpha-1}$. [Preuve dans (Araya-López et al.*, 2010, annexe)]*

Sous ces conditions, on peut montrer que l'approximation PWLC est bornée.

Theorème 4
Soit ρ une fonction continue et convexe sur Δ, dérivable partout sur Δ^o (l'intérieur de Δ), et satisfaisant la condition α-Hölder avec la constante K_α. L'erreur d'une approximation ω_B est majorée par ${C\delta_B}^\alpha$, où C est une constante scalaire. [Preuve dans (Araya-López et al.*, 2010, annexe)]*

4.2. Mises à jour exactes

Sachant que l'approximation de ρ est bornée pour une large famille de fonctions, les techniques décrites en section 3.3.1. peuvent être directement appliquées en utilisant $\omega_B(b)$ comme fonction de récompense PWLC. Ces algorithmes peuvent être employés sans danger parce que la propagation de l'erreur liée aux mises à jour exactes est bornée. Cela peut être prouvé en utilisant une méthode similaire à (Pineau *et al.*, 2006; Lovejoy, 1991). Soit V_t la fonction de valeur utilisant l'approximation PWLC décrite ci-dessus et V_t^* la fonction de valeur optimale, toutes les deux au temps t, H étant l'opérateur de mise à jour exact et $\hat{H}$ le même opérateur avec l'approximation PWLC. Alors l'erreur avec la fonction de valeur réelle est

$$
\begin{aligned}
\|V_t - V_t^*\|_\infty &= \|\hat{H}V_{t-1} - HV_{t-1}^*\|_\infty && \text{(par definition)} \\
&\leq \|\hat{H}V_{t-1} - HV_{t-1}\|_\infty + \|HV_{t-1} - HV_{t-1}^*\|_\infty \\
& && \text{(par inégalité triangulaire)} \\
&\leq |\omega_{b^*} + \alpha_{b^*} \cdot b - \rho(b) - \alpha_{b^*} \cdot b| + \|HV_{t-1} - HV_{t-1}^*\|_\infty \\
& && \text{(erreur maximale en } b\text{)} \\
&\leq {C\delta_B}^\alpha + \|HV_{t-1} - HV_{t-1}^*\|_\infty && \text{(par théorème 4)} \\
&\leq {C\delta_B}^\alpha + \gamma\|V_{t-1} - V_{t-1}^*\| && \text{(par contraction)} \\
&\leq \frac{{C\delta_B}^\alpha}{1-\gamma} && \text{(par somme d'une série géométrique)}
\end{aligned}
$$

Pour ces algorithmes, la sélection de l'ensemble B reste ouverte, soulevant les mêmes questions que pour la sélection des états de croyance pour les algorithmes PB.

4.3. Mises à jour approchées

Dans le cas des algorithmes PB, l'extension est directe elle aussi, et les algorithmes décrits en section 3.3.2. peuvent être utilisés avec une erreur bornée. La sélection de B, l'ensemble des points de l'approximation PWLC, et la sélection de l'ensemble des points pour l'algorithme peuvent être partagées. [5] Cela simplifie l'étude de la majoration quand on utilise les deux techniques d'approximation en même temps. Soit $\hat{V}_t$ la fonction de valeur au temps t calculée en utilisant l'approximation PWLC et un algorithme PB. Alors l'erreur entre $\hat{V}_t$ et V_t^* est $\|\hat{V}_t - V_t^*\|_\infty \leq \|\hat{V}_t - V_t\|_\infty + \|V_t - V_t^*\|_\infty$. Le second terme est le même que dans la section 4.2., il est donc majoré par $\frac{C{\delta_B}^\alpha}{1-\gamma}$. Le premier terme peut être majoré en suivant le même raisonnement que dans (Pineau *et al.*, 2006), où $\|\hat{V}_t - V_t\|_\infty \leq \frac{(R_{max}-R_{min}+C{\delta_B}^\alpha)\delta_B}{1-\gamma}$, avec R_{min} et R_{max} les valeurs respectivement minimale et maximale de $\rho(b)$. La raison est que le pire cas pour un vecteur α est $\frac{R_{min}-\epsilon}{1-\gamma}$, alors que le meilleur cas est seulement $\frac{R_{max}}{1-\gamma}$ parce que l'approximation est toujours une minoration.

5. Conclusions

Nous avons introduit les ρPOMDP, une extension des POMDP qui permet d'exprimer des problèmes de prise de décision séquentielle dans lesquels réduire l'incertitude sur certaines variables d'état est un objectif explicite. Dans ce modèle, la récompense ρ est typiquement une fonction convexe de l'état de croyance.

En utilisant la convexité de ρ, un premier résultat important que nous prouvons est que la mise à jour de Bellman $V_n = HV_{n-1}$ préserve la convexité. En particulier, si ρ est PWLC et que la fonction de valeur V_0 est égale à 0, alors V_n est aussi PWLC et il est simple d'adapter de nombreux algorithmes de résolution de POMDP de l'état de l'art. Nous proposons pour cela d'employer des approximations PWLC des fonctions de récompense convexes pour nous ramener à un cas plus simple, et montrons que les algorithmes obtenus convergent à la limite vers la fonction de valeur optimale.

Des travaux précédents ont déjà introduit des récompenses dépendant de la croyance, comme dans la discussion de Spaan sur les POMDP et la perception active (Spaan, 2008), ou le travail de Hero *et al.* sur la gestion de capteurs avec des POMDP (Hero *et al.*, 2007). Cependant, le premier ne fait

5. Les points de la frontière de Δ doivent être retirés puisque les preuves ne reposent que sur des points intérieurs.

que présenter le problème des fonctions de valeurs non-PWLC sans fournir de solution spécifique, alors que le second résout le problème en employant des techniques de Monte-Carlo qui ne se basent pas sur la propriété PWLC. Dans le domaine de la robotique, des mesures d'incertitude dans des POMDP ont été largement utilisées comme heuristiques (Thrun, 2000), avec de très bons résultats mais sans garanties de convergence. Ces techniques n'utilisent que des récompenses dépendant de l'état, mais des mesures d'incertitude sont employées pour accélérer le processus de résolution, au prix de perdre quelques propriétés élémentaires (par exemple la propriété de Markov). Notre travail ouvre la voie à la résolution de problèmes avec récompense dépendant de la croyance en utilisant de nouveaux algorithmes d'approximation de la fonction de valeur (par exemple à base de points) d'une manière théoriquement aboutie.

Un point important est que la complexité temporelle de ces nouveaux algorithmes change du fait de la taille de l'approximation de ρ. Les travaux futurs incluent la réalisation d'expérimentations pour mesurer l'augmentation de cette complexité. Une tâche plus complexe est d'évaluer la qualité des approximations résultantes du fait du manque d'autres algorithmes pour ρPOMDP. Une possibilité est de considérer les algorithmes de Monte-Carlo en ligne (Ross *et al.*, 2008), lesquels devraient requérir peu de modifications.

Remerciements

Ces travaux de recherche ont eu le soutien d'un financement doctoral *CONICYT-Ambassade de France* et du projet COMAC. Nous souhaitons aussi remercier Bruno Scherrer pour les discussions fructueuses et les relecteurs anonymes pour leurs commentaires et suggestions utiles.

Références

ARAYA-LÓPEZ M., BUFFET O., THOMAS V. & CHARPILLET F. (2010). *A POMDP Extension with Belief-dependent Rewards – Extended Version*. Rapport interne RR-7433, INRIA. (See also NIPS supplementary material).

BELLMAN R. (1954). The theory of dynamic programming. *Bull. Amer. Math. Soc.*, **60**, 503–516.

CASSANDRA A. (1998). *Exact and approximate algorithms for partially observable Markov decision processes*. PhD thesis, Brown University, Providence, RI, USA.

COVER T. & THOMAS J. (1991). *Elements of Information Theory*. Wiley-Interscience.

HAUSKRECHT M. (2000). Value-function approximations for partially observable Markov decision processes. *Journal of Artificial Intelligence Research*, **13**, 33–94.

HERO A., CASTAN D., COCHRAN D. & KASTELLA K. (2007). *Foundations and Applications of Sensor Management*. Springer Publishing Company, Incorporated.

JI S. & CARIN L. (2007). Cost-sensitive feature acquisition and classification. *Pattern Recogn.*, **40**(5), 1474–1485.

KAPLOW R. (2010). Point-based POMDP solvers : Survey and comparative analysis. Master's thesis, McGill University, Montreal, Quebec, Canada.

KURNIAWATI H., HSU D. & LEE W. (2008). SARSOP : Efficient point-based POMDP planning by approximating optimally reachable belief spaces. In *Robotics : Science and Systems IV*.

LOVEJOY W. (1991). Computationally feasible bounds for partially observed Markov decision processes. *Operations Research*, **39**(1), 162–175.

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

MONAHAN G. (1982). A survey of partially observable Markov decision processes. *Management Science*, **28**, 1–16.

PINEAU J., GORDON G. & THRUN S. (2006). Anytime point-based approximations for large POMDPs. *Journal of Artificial Intelligence Research (JAIR)*, **27**, 335–380.

ROSS S., PINEAU J., PAQUET S. & CHAIB-DRAA B. (2008). Online planning algorithms for POMDPs. *Journal of Artificial Intelligence Research (JAIR)*, **32**, 663–704.

SAIGAL R. (1979). On piecewise linear approximations to smooth mappings. *Mathematics of Operations Research*, **4**(2), 153–161.

SMALLWOOD R. & SONDIK E. (1973). The optimal control of partially observable Markov decision processes over a finite horizon. *Operation Research*, **21**, 1071–1088.

SMITH T. & SIMMONS R. (2005). Point-based POMDP algorithms : Improved analysis and implementation. In *Proc. of the Int. Conf. on Uncertainty in Artificial Intelligence (UAI)*.

SPAAN M. (2008). Cooperative active perception using POMDPs. In *AAAI 2008 Workshop on Advancements in POMDP Solvers*.

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

THRUN S. (2000). Probabilistic algorithms in robotics. *AI Magazine*, **21**(4), 93–109.