MDP s-lipschitziens et ρ-POMDP non-convexes

Olivier Buffet[1,2], Vincent Thomas[2,1], Jilles Dibangoye[3]

[1] INRIA Nancy Grand-Est, LORIA
615, rue du jardin botanique, Villers-lès-Nancy
prenom.nom@loria.fr et https://members.loria.fr/olivier.buffet

[2] Université de Lorraine / CNRS, LORIA
Campus scientifique Victor Grignard, Vandœuvre-lès-Nancy
prenom.nom@loria.fr et https://members.loria.fr/vincent.thomas

[3] Univ Lyon, INSA Lyon, Inria, CITI
6 Avenue des Arts, F-69621 Villeurbanne
prenom.nom@inria.fr et http://dibangoye.fr

Résumé : Cet article s'intéresse aux MDP à états continus quand leurs fonctions de transition et de récompense sont lipschitziennes (par rapport à l'état). Il montre que, dans ce cadre, la fonction de valeur optimale à horizon fini est elle-même lipschitzienne, ce qui permet d'employer des représentations adaptées (en dents de scie : vers le bas pour un majorant et vers le haut pour un minorant) et de proposer des algorithmes de résolution à erreur bornée directement inspirés d'algorithmes de l'état de l'art. Cela permettra en particulier de résoudre des ρPOMDP (Araya-López *et al.*, 2010) – c'est-à-dire des POMDP dont la récompense dépend de l'état de croyance – même si la fonction de récompense n'est pas convexe dans l'espace des états de croyance. Un objectif à plus long terme est de pouvoir adopter des méthodes similaires pour résoudre des jeux de Markov partiellement observables (POSG). **Mots-clés** : MDP, état continu, POMDP, ρ-POMDP, Lipschitz, HSVI

1 Introduction

Nous nous intéressons au moyen de résoudre des MDP à états continus quand leurs fonctions de transition et de récompense sont relativement régulières (*smooth*), plus précisément quand elles sont lipschitziennes (par rapport à l'état). Cette situation se rencontre dans tout belief MDP utilisé dans la résolution d'un POMDP, cas dans lequel on préférera exploiter la convexité de la fonction de valeur. Dans le cas des ρPOMDP (Araya-López *et al.*, 2010) – c'est-à-dire des POMDP dont la récompense dépend de l'état de croyance – la convexité de la fonction de valeur n'est garantie que si la fonction de récompense est elle-même convexe. Par ailleurs, on pourra envisager dans certains cas d'approcher un problème à l'aide d'un modèle lipschitzien.

Comme on va le voir, sous l'hypothèse que la dynamique comme la fonction de récompense sont lipschitziennes (par rapport à l'état), la fonction de valeur optimale à horizon fini est, elle aussi, lipschitzienne, ce qui permet de l'approcher ou de l'encadrer en contrôlant l'erreur faite à l'aide d'approximateurs "en dents-de-scie coniques". On va ainsi pouvoir proposer des algorithmes de résolution à horizon temporel fini ou infini, et avec ou sans état initial. Dans chacun de ces cas, on suivra une démarche similaire à celle suivie dans de multiples solveurs de POMDP (exploitant, eux, la convexité de la fonction de valeur), mais en utilisant des approximateurs non-convexes.

Les sections 2 et 3 présentent respectivement des travaux connexes et un bref état de l'art sur les MDP, POMDP et ρPOMDP. La section 4 décrit les résultats de lipschitz-continuité de la fonction de valeur optimale obtenus, en proposant essentiellement des majorants de la constante de Lipschitz. Sur cette base, la section 5 explique comment approcher la fonction de valeur optimale V^* en exploitant cette propriété de Lipschitz, et comment adapter des algorithmes de l'état de l'art dans cette situation, tout en contrôlant l'erreur faite. Une discussion vient conclure sur les perspectives ouvertes par cette approche.

2 Travaux connexes

L'exploitation de la propriété de linéarité par morceaux et de convexité (PWLC) de la fonction de valeur dans les POMDP (Sondik, 1971; Smallwood & Sondik, 1973) est très comparable à l'exploitation de la Lipschitz-continuité que nous allons aborder ici. Cette dernière est une hypothèse moins forte, donc applicable dans un cadre plus large. Elle fournit des approximateurs moins efficaces (nécessitant une résolution plus fine pour obtenir la même précision), mais qui permettront quand même de contrôler l'erreur faite.

Plus proches du présent travail concernant la Lipschitz-continuité du problème, Ieong *et al.* (2007) ont considéré un cadre reposant sur les hypothèses suivantes :

— la dynamique du MDP est déterministe (ce qui ne sera qu'un cas particulier pour nous) ;
— l'espace des états et l'espace des actions sont continus ;
— le modèle de la dynamique (T) et le modèle de récompense (r) sont lipschitziens par rapport aux états comme aux actions ;
— une heuristique admissible lipschitzienne est disponible ;
— l'objectif est de trouver un chemin le plus court jusqu'à un état but en présence de culs-de-sac.

Sur cette base, ils proposent un algorithme de type "meilleur d'abord", lequel doit subir d'onéreuses mises à jour de ses minorants (dans un cadre de minimisation). En comparaison, nous (i) ignorons les actions continues, (ii) tenons compte de dynamiques stochastiques, (iii) nous attaquons à des problèmes à horizon fini ou infini, (iv) caractérisons la forme de la fonction de valeur, et (v) nous orientons plutôt vers des algorithmes reposant sur des générations de trajectoires. Comme argumenté par Smith & Simmons (2004), (a) de tels algorithmes en "profondeur d'abord" permettent un meilleur compromis entre consommation mémoire et temps de calcul, et (b) des algorithmes de type "meilleur d'abord" (comme celui de Ieong *et al.* (2007)) doivent propager des mises à jours des bornes dans leur file des priorités, ce qui est très coûteux.

3 Etat de l'art

Nous allons aborder deux cadres : d'une part celui des MDP (à espace d'état continu) dits s-lipschitziens, et d'autre part celui des ρ-POMDP. Comme on pourra l'observer, les résultats théoriques sont proches dans les deux cas.

3.1 MDP s-lipschitziens

Un MDP (Bellman, 1957) est ici défini par un tuple $\langle \mathcal{S}, \mathcal{A}, T, r \rangle$ dans lequel :

— $\mathcal{S}$ est un ensemble continu d'états, doté d'une métrique $d(\cdot, \cdot')$;
— $\mathcal{A}$ est un ensemble fini d'actions ;
— T est une fonction de transition soit déterministe (de sorte qu'appliquer une action a dans un état s amène toujours au même état $s' = T(s, a)$), soit stochastique (de sorte qu'appliquer une action a dans un état s amène à un état s échantillonné selon une distribution $T(s, a)$) ; et
— r est une fonction de récompense définie sur $\mathcal{S} \times \mathcal{A}$ et à valeurs dans $\mathbb{R}$; on supposera en outre r majorée (respectivement minorée) par une constante R^{max} (resp. R^{min}).

Dans le cas déterministe, le caractère s-lipschitzien des MDP vient du fait que les fonctions de transition et de récompense sont lipschitziennes par rapport à l'état, ce qui s'écrit, avec les constantes ρ et τ (pour tout s, s', a) :

$$|r(s, a) - r(s', a)| \leq \rho d(s, s'),$$
$$d\left(T(s, a), T(s', a)\right) \leq \tau d(s, s'),$$

où $d(\cdot, \cdot)$ est la métrique sur $\mathcal{S}$.

Dans le cas stochastique, il n'est pas possible de traiter la fonction de transition de la même manière, essentiellement parce qu'il est difficile de définir une métrique sur des distributions de probabilité. Nous faisons ici l'hypothèse (inspirée indirectement par Somani *et al.* (2013)) qu'il est possible d'écrire la dynamique stochastique du système avec une fonction de transition déterministe

$$T : \mathcal{S} \times \mathcal{A} \times \mathcal{X} \to \mathcal{S},$$

où $\mathcal{X}$ est le domaine d'une variable aléatoire X suivant une densité de probabilité conditionnelle $f(x|s, a)$. En d'autres termes, (i) la variable aléatoire X représente un événement dont la probabilité d'occurrence

dépend de s et a, et (ii) cet événement a une influence sur le prochain état qui dépend aussi de s et a. Nous verrons que ces deux propriétés sont aussi présentes dans les POMDP. Avec cette modélisation, le caractère s-lipschitzien de la fonction de transition passe par ces deux étapes ($\forall a \in \mathcal{A},\ x \in \mathcal{X},\ (s,s') \in \mathcal{S}^2$) :

$$|f(x|s,a) - f(x|s',a)| \leq \phi d(s,s'), \text{ et}$$
$$d\left(T(s,a,x), T(s',a,x)\right) \leq \tau d(s,s').$$

Dans tous les cas (déterministe ou stochastique), on cherche ici des politiques optimisant un critère total avec facteur d'atténuation γ et horizon temporel (éventuellement infini) H :

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\pi}\left[\sum_{t=0}^{H} \gamma^t R_t\right].$$

On utilisera pour cela des algorithmes calculant tout ou partie de la fonction de valeur optimale obtenue à l'aide de l'opérateur de Bellman ($\forall s \in \mathcal{S}$) :

$$V_t^*(s) = 0, \text{ et}$$
$$V_{t-1}^*(s) = \max_a \left[r(s,a) + \gamma \int_{s'} P(s'|s,a) V_t^*(s') ds'\right] \quad (\forall t \in \{0,\dots,H-1\}).$$

Cette fonction de valeur sera couramment majorée par

$$V_t^{max} = \begin{cases} \frac{1-\gamma^{H-t}}{1-\gamma} R^{max} & \text{si } \gamma < 1; \\ (H-t)R^{max} & \text{si } \gamma = 1; \end{cases} \qquad (\forall t \in \{0,\dots,H\})$$

voire, pour éviter la dépendance en t,

$$V^{max} = \begin{cases} \frac{R^{max}}{1-\gamma} & \text{si } \gamma < 1; \\ HR^{max} & \text{si } \gamma = 1. \end{cases}$$

On définira de même les minorants V_t^{min} et V^{min}, mais aussi $V_t^{lim} = \max\{|V_t^{max}|, |V_t^{min}|\}$ et $V^{lim} = \max\{|V^{max}|, |V^{min}|\}$.

3.2 POMDP et ρ-POMDP

Nous rappelons d'abord la définition d'un POMDP avant de passer à celle, moins connue, d'un ρ-POMDP.

Un MDP partiellement observable (POMDP) (Astrom, 1965) est défini par un tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{Z}, T, O, r\rangle$ où

— $\langle \mathcal{S}, \mathcal{A}, T, r\rangle$ est un MDP ;
— $\mathcal{Z}$ est un ensemble fini d'observations possibles ; et
— O est une fonction d'observation donnant $O(a,s',z) = Pr(z|a,s')$, la probabilité d'observer z si l'action a conduit dans l'état s'.

La recherche d'une politique optimale se fait souvent en passant par le MDP sur les états de croyance (bMDP) correspondant, dans lequel

— la connaissance actuelle de la situation est résumée par un état de croyance b, c'est-à-dire une distribution de probabilité sur les états possibles ;
— la fonction de transition $T(b,a,b')$ donne la probabilité d'aller d'un état de croyance b à un autre b' étant donnée une action a (en tenant compte des observations qui pourront être rencontrées) ; et
— la fonction de récompense donne la récompense moyenne étant donnée l'état de croyance : $r(b,a) = \sum_{s\in\mathcal{S}} b(s) r(s,a)$.

Pour des raisons pratiques, nous emploierons les mêmes notations T et r qu'il s'agisse des fonctions décrivant un MDP, un POMDP, ou son bMDP associé.

Araya-López *et al.* (2010) ont étendu le cadre des POMDP pour pouvoir traiter non seulement des problèmes de contrôle de l'état, mais aussi de recherche d'information. Pour cela, ils remplacent la donnée de $r(s,a)$ par une fonction de récompense définie sur l'état de croyance $r(b)$, par exemple liée à une mesure d'entropie. [1] Pour des raisons pratiques, nous ne notons pas cette fonction ρ comme les auteurs, réservant ρ pour dénoter la constante de Lipschitz associée à cette fonction le cas échéant. Dans ce contexte, Araya-López *et al.* ont montré que :

1. On notera que cette fonction de récompense peut aussi dépendre de s et/ou a.

— si r est linéaire par morceaux et convexe (PWLC) sur $\mathcal{B}$, le simplexe des états de croyance, alors la fonction de valeur est aussi PWLC, et de nombreux solveurs reposant sur cette propriété peuvent être exploités ;
— si r est convexe et λ-lipschitzienne ou α-hölderienne, alors on peut l'approcher arbitrairement bien par une fonction PWLC, et donc approcher la fonction de valeur arbitrairement bien.

Dans cet article, nous allons montrer comment se passer de la contrainte de convexité de r, tout en mettant en œuvre une approche assez similaire.

4 Lipschitz-continuité de V^*

Avant de présenter les résultats de Lipschitz-continuité de V^* dans différents cadres, la section suivante décrit quelques résultats préliminaires utiles.

4.1 Résultats préliminaires

Du fait de l'opérateur de Bellman utilisé pour les mises à jour de la fonction de valeur, dans les différents scénarios que nous allons étudier, nous rencontrerons des suites arithmético-géométriques, en l'occurrence de la forme $u_t = \alpha \cdot u_{t+1} + \beta$ initialisée par $u_H = 0$, pour t allant de H à 0, et avec $\alpha > 0$. On a alors le résultat suivant.

Lemme 1
On a, pour tout $t \in \{0, \cdots, H\}$,

$$u_t = \begin{cases} \frac{1-\alpha^{H-t}}{1-\alpha}\beta & \text{si } \alpha \neq 1, \text{ et} \\ (H-t)\beta & \text{si } \alpha = 1. \end{cases}$$

Démonstration 1
La preuve est immédiate en modifiant les indices de résultats classiques pour les suites arithmético-géométriques (et avec une initialisation à 0*).* [2] □

Par ailleurs, nous aurons besoin de la propriété suivante pour le produit de deux fonctions scalaires lipschitziennes.

Lemme 2
Soit X *un ensemble doté d'une métrique* d*, et soient* f *et* g *deux fonctions de* X *dans* $\mathbb{R}$*, lipschitziennes (de constantes respectives* k_f *et* k_g*). Alors, pour tout* $(x, x') \in X^2$,

$$|f(x) \cdot g(x) - f(x') \cdot g(x')| \leq (|g(x)| \cdot k_f + |f(x)| \cdot k_g) \cdot d(x, x').$$

Démonstration 2
Pour tout $(x, x') \in X^2$,

$$\begin{aligned} |f(x) \cdot g(x) - f(x') \cdot g(x')| &= |f(x) \cdot g(x) - f(x') \cdot g(x) + f(x') \cdot g(x) - f(x') \cdot g(x')| \\ &\leq |f(x) \cdot g(x) - f(x') \cdot g(x)| + |f(x') \cdot g(x) - f(x') \cdot g(x')| \\ &\leq |g(x) \cdot (f(x) - f(x'))| + |f(x') \cdot (g(x) - g(x'))| \\ &\leq |g(x)| \cdot |f(x) - f(x')| + |f(x')| \cdot |g(x) - g(x')| \\ &\leq |g(x)| \cdot k_f \cdot d(x, x') + |f(x)| \cdot k_g \cdot d(x, x') \\ &= (|g(x)| \cdot k_f + |f(x)| \cdot k_g) \cdot d(x, x'). \end{aligned}$$

□

En outre, si $|f|$ et $|g|$ sont majorées respectivement par F et $G \in \mathbb{R}$, alors

$$|f(x) \cdot g(x) - f(x') \cdot g(x')| \leq \underbrace{(G \cdot k_f + F \cdot k_g)}_{k_{f \cdot g}} \cdot d(x, x'),$$

donc la fonction $f \cdot g$ est lipschitzienne.

2. https://fr.wikipedia.org/wiki/Suite_arithm\%C3\%A9tico-g\%C3\%A9om\%C3\%A9trique

4.2 MDP s-lipschitziens, cas déterministe

Dans le cas des MDP s-lipschitziens à dynamique déterministe, on a le résultat suivant.

Lemme 3
Sur un horizon fini H, la fonction de valeur optimale vérifie, $\forall t \in \{0, \dots, H\}$, $\forall s, s' \in S$,

$$|V_t^*(s) - V_t^*(s')| \leq \nu_t^{max} d(s, s'),$$
$$\text{où } \nu_t^{max} = \begin{cases} \rho \frac{1-(\gamma\tau)^{H-t}}{1-\gamma\tau} & \text{si } \gamma\tau \neq 1, \\ \rho(H-t) & \text{si } \gamma\tau = 1. \end{cases}$$

Démonstration 3
Pour $t = H$, on a, pour tout (s, s'), $|V_H^*(s) - V_H^*(s')| = |0 - 0| = 0$, donc la propriété est vérifiée.
Supposons maintenant que la propriété est vérifiée pour $t \in \{1, \dots, H\}$. Alors, pour tout (s, s'),

$$\begin{aligned} V_{t-1}^*(s) &= \max_a \left[r(s,a) + \gamma V_t^*(T(s,a)) \right] \\ &\leq \max_a \Big[\underbrace{\left(r(s',a) + \rho d(s,s') \right)}_{r \text{ lipschitz.}} + \gamma \underbrace{\left(V_t^*(T(s',a)) + \nu_t^{max} \tau d(s,s') \right)}_{V_t^* \text{ et } T \text{ lipschitz.}} \Big] \\ &= \underbrace{\max_a \left[r(s',a) + \gamma V_t^*(T(s',a)) \right]}_{V_{t-1}^*(s')} + \left(\rho d(s,s') + \gamma \nu_t^{max} \tau d(s,s') \right), \end{aligned}$$

d'où

$$V_{t-1}^*(s) - V_{t-1}^*(s') \leq \rho d(s,s') + \gamma \nu_t^{max} \tau d(s,s') = \underbrace{\left(\gamma\tau \nu_t^{max} + \rho \right)}_{\nu_{t-1}^{max}} d(s,s').$$

On peut alors appliquer le lemme 1 avec $u_t = \nu_t^{max}$, $\alpha = \gamma\tau$ et $\beta = \rho$ pour obtenir le résultat escompté pour tout t. □

Notons que, quand l'horizon H croît, la constante de Lipschitz pour $t = 0$ ainsi obtenue tend vers :

$$\begin{cases} \rho \frac{1}{1-\gamma\tau} & \text{si } \gamma\tau < 1, \\ +\infty & \text{sinon (linéairement ssi } \gamma\tau = 1). \end{cases}$$

Autrement dit, le facteur d'atténuation γ peut compenser l'expansion de l'espace d'état due à la dynamique du système. A horizon infini, si $\gamma\tau \geq 1$, la fonction de valeur optimale n'est pas nécessairement lipschitzienne, mais il est possible de se ramener à un horizon fini en cherchant une solution ϵ-optimale. A ce propos, on remarquera que, même dans des problèmes de chemin le plus court déterministes, le nombre minimum de pas de temps pour atteindre le but n'est pas majoré, à moins que :

— l'ensemble des états soit fini (donc non continu), ou
— l'ensemble des actions garantisse que l'ensemble des états *atteignables* soit fini.

Cela pourrait expliquer que Ieong *et al.* (2007) ne font pas l'hypothèse que la fonction de valeur est lipschitzienne et ne cherchent pas à majorer la constante associée, mais se reposent sur l'existence d'une fonction heuristique (admissible) qui soit lipschitzienne. [3]

4.3 MDP s-lipschitziens, cas stochastique

Dans le cas des MDP s-lipschitziens à dynamique stochastique, des constantes de Lipschitz peuvent être calculées de manière récursive, comme décrit par le lemme suivant.

Lemme 4
Supposons que $\mathcal{X}$ soit borné, donc ait un volume fini $vol(\mathcal{X})$. Alors, sur un horizon fini H, la fonction de valeur optimale est ν_t-lipschitzienne pour tout $t \in \{0, \dots, H\}$, où

$$\nu_H = 0, \quad \text{et, pour } t \in \{1, \dots, H\},$$
$$\nu_{t-1} = (\gamma\tau)\nu_t + \left(\rho + \gamma V_t^{lim} \phi vol(\mathcal{X}) \right).$$

3. On observera qu'il n'est pas nécessaire que la fonction de valeur optimale soit lipschitzienne pour qu'elle puisse être minorée ou majorée par une fonction lipschitzienne.

Démonstration 4
Pour $t = H$, on a, pour tout (s, s'), $|V^*_H(s) - V^*_H(s')| = |0 - 0| = 0$, donc la propriété est vérifiée.
Supposons que la propriété soit vérifiée pour $t \in \{1, \dots, H\}$. Alors, pour tout (s, s'),

$$V^*_{t-1}(s) = \max_a \left[r(s,a) + \gamma \int_{x \in \mathcal{X}} V^*_t(T(s,a,x)) f(x|s,a) dx \right]$$

Or, V_t, T, et f étant lipschitziennes, en appliquant le lemme 2 et en utilisant le fait que la composition de fonctions lipschitziennes est lipschitzienne,

$$\begin{aligned}
&V^*_t(T(s,a,x)) f(x|s,a) \\
&\leq V^*_t(T(s',a,x)) f(x|s',a) + \Big(\underbrace{|f(x|s',a)|}_{\leq 1 \text{ après intégration}} \cdot \nu_t \tau + \underbrace{|V^*_t(T(s',a,x))|}_{\leq V^{lim}_t} \cdot \phi \Big) \cdot d(s,s').
\end{aligned}$$

D'où,

$$\begin{aligned}
&V^*_{t-1}(s) \\
&\leq \max_a \Big[\underbrace{(r(s',a) + \rho d(s,s'))}_{r \text{ lipschitz.}} \\
&\qquad + \gamma \int_{x \in \mathcal{X}} \left[V^*_t(T(s',a,x)) f(x|s',a) + \left(f(x|s',a) \cdot \nu_t \tau + V^{lim}_t \cdot \phi \right) \cdot d(s,s') \right] dx \Big] \\
&\leq \max_a \Big[(r(s',a) + \rho d(s,s')) + \gamma \Big[\int_{x \in \mathcal{X}} V^*_t(T(s',a,x)) f(x|s',a) dx \\
&\qquad + \nu_t \tau d(s,s') \underbrace{\int_{x \in \mathcal{X}} f(x|s',a) dx}_{=1} + V^{lim}_t \phi d(s,s') \underbrace{\int_{x \in \mathcal{X}} 1 dx}_{=vol(\mathcal{X})} \Big] \Big] \\
&= \max_a \left[r(s',a) + \gamma \int_{x \in \mathcal{X}} V^*_t(T(s',a,x)) f(x|s',a) dx \right] \\
&\qquad + \left(\rho d(s,s') + \gamma \left[\nu_t \tau d(s,s') + V^{lim}_t \phi d(s,s') vol(\mathcal{X}) \right] \right) \\
&= V^*_{t-1}(s') + \left(\rho + \gamma \left[\tau \nu_t + V^{lim}_t \phi vol(\mathcal{X}) \right] \right) d(s,s') \\
&= V^*_{t-1}(s') + \underbrace{\left((\gamma\tau)\nu_t + \left(\rho + \gamma V^{lim}_t \phi vol(\mathcal{X}) \right) \right)}_{\nu_{t-1}} d(s,s').
\end{aligned}$$

On a donc bien la formule de récurrence attendue. □

Cette suite de constantes de Lipschitz est exploitable dans les algorithmes, mais il est difficile de déterminer leur évolution quand l'horizon croît. En faisant des approximations plus grossières, on peut majorer cette suite par une autre suite de constantes de Lipschitz dont on peut obtenir une expression non récursive, comme détaillé dans le lemme suivant.

Lemme 5
Supposons que $\mathcal{X}$ soit borné, donc ait un volume fini $vol(\mathcal{X})$. Alors, sur un horizon fini H, la fonction de valeur optimale vérifie, $\forall t \in \{0, \dots, H\}$, $\forall s, s' \in S$,

$$|V^*_t(s) - V^*_t(s')| \leq \nu^{max}_t d(s,s'),$$

$$\text{où } \nu^{max}_t = \begin{cases} \left(\rho + \gamma V^{lim} \phi vol(\mathcal{X}) \right) \dfrac{1 - (\gamma\tau)^{H-t}}{1 - \gamma\tau} & \text{si } \gamma\tau \neq 1, \\ \left(\rho + \gamma V^{lim} \phi vol(\mathcal{X}) \right) (H - t) & \text{si } \gamma\tau = 1. \end{cases}$$

Démonstration 5
Comme précédemment, la propriété est trivialement vérifiée pour $t = H$.
En remplaçant, pour tout $t \in \{1, \dots, H\}$, V^{lim}_t par $V^{lim} (> V^{lim}_t)$ dans l'expression reliant ν_{t-1} à ν_t, on construit une suite $(\nu^{max}_t)_{t \in \{0,\dots,H\}}$ vérifiant, pour tout $t \in \{1, \dots, H\}$,

$$\nu^{max}_{t-1} = \underbrace{(\gamma\tau)}_{\alpha} \nu^{max}_t + \underbrace{\left(\rho + \gamma V^{lim} \phi vol(\mathcal{X}) \right)}_{\beta}.$$

On peut alors appliquer le lemme 1 pour obtenir le résultat escompté pour tout t. □

On peut remarquer à propos de la suite de constantes de Lipschitz ainsi obtenue :
— qu'elle diverge ici dès que $\gamma\tau \geq 1$;
— qu'en prenant $\phi = 0$ (indépendance de la première "étape" par rapport à s), on retrouve l'expression obtenue dans le cas déterministe) ;
— qu'elle dépend du volume (fini) de $\mathcal{X}$; et
— que $vol(\mathcal{X})$ et ϕ sont liés puisqu'on pourrait ré-écrire le même modèle de dynamique en effectuant des changements d'échelle ; on pourrait ainsi toujours se ramener à $vol(\mathcal{X}) = 1$.

4.4 POMDP et ρ-POMDP

Partant du fait que la fonction de valeur d'un POMDP à horizon fini est linéaire par morceaux et convexe (PWLC), il est évident qu'elle est aussi lipschitzienne. Nous allons ici encadrer la constante de Lipschitz associée à chaque horizon, et ce dans le cas plus général des ρ-POMDP avec une fonction de récompense r qui est lipschitzienne de constante ρ.

Comme précédemment, on va chercher à encadrer l'écart entre la fonction de valeur optimale en un état de croyance b_1 et un autre b_2. Nous allons pour cela avoir besoin des résultats préliminaires suivants. [4]

Lemme 6
Etant donnés deux états de croyance b_1 et b_2, une action effectuée a et une observation reçue z, on a :

$$\begin{aligned}\|b_1^{a,z} - b_2^{a,z}\|_\infty &\leq \lambda_{a,z}\|b_1 - b_2\|_\infty \\ \text{où } \lambda_{a,z} &= \max_{s'} \sum_s \frac{P(z|a,s')P(s'|s,a)}{\sum_{s"} P(z|a,s")P(s"|s,a)} \\ &= \max_{s'} \sum_s \frac{P(z,s'|s,a)}{\sum_{s"} P(z,s"|s,a)} \\ &= \max_{s'} \sum_s P(s'|s,a,z).\end{aligned}$$

Démonstration 6
Rappelons d'abord le calcul de la mise à jour d'un état de croyance pour une paire action-observation (a,z) :

$$\begin{aligned}\forall s', \quad b^{a,z}(s') = P(s'|b,a,z) &= \sum_s P(s'|s,a,z)b(s) \\ &= \sum_s \frac{P(z|a,s')P(s'|s,a)}{\sum_{s"} P(z|a,s")P(s"|s,a)} b(s).\end{aligned}$$

De là, on tire :

$$\begin{aligned}\forall s', \quad b_1^{a,z}(s') - b_2^{a,z}(s') &= \sum_s \frac{P(z|a,s')P(s'|s,a)}{\sum_{s"} P(z|a,s")P(s"|s,a)} \left[b_1(s) - b_2(s)\right] \\ &\leq \left[\sum_s \frac{P(z|a,s')P(s'|s,a)}{\sum_{s"} P(z|a,s")P(s"|s,a)}\right] \|b_1 - b_2\|_\infty \\ &\leq \max_{s'} \left[\sum_s \frac{P(z|a,s')P(s'|s,a)}{\sum_{s"} P(z|a,s")P(s"|s,a)}\right] \|b_1 - b_2\|_\infty.\end{aligned}$$

D'où le résultat attendu. □

En pratique, on utilisera comme constante de Lipschitz $\lambda = \max_{a,z} \lambda_{a,z}$.

4. Dans le cadre des POMDP, Platzman (1977) montre un résultat équivalent au lemme 6, mais pour une distance entre états de croyance différente, et avec une constante de Lipschitz inférieure ou égale à 1. Une question ouverte est donc de savoir si, en démontrant le même résultat pour la norme infinie ou en utilisant la même distance que Platzman, on ne pourrait pas obtenir de meilleurs encadrements de la fonction de valeur optimale que dans le présent article.

Lemme 7
Etant donnés deux états de croyance b_1 et b_2, une action effectuée a et une observation reçue z, on a :

$$|P(z|b_1,a) - P(z|b_2,a)| \leq \mu_{a,z}\|b_1 - b_2\|_\infty$$
$$\text{où} \quad \mu_{a,z} = \sum_s P(z|a,s) = \sum_{s,s'} P(z|a,s')T(s,a,s').$$

Démonstration 7
Le calcul de la probabilité d'une observation z pour un état de croyance b et une action a donne :

$$P(z|b,a) = \sum_s P(z|s,a)b(s) = \sum_{s,s'} P(z|a,s')T(s,a,s')b(s).$$

De là, on tire :

$$\begin{aligned} P(z|b_1,a) - P(z|b_2,a) &= \sum_{s,s'} P(z|a,s')T(s,a,s')\left[b_1(s) - b_2(s)\right] \\ &\leq \Big[\sum_{s,s'} P(z|a,s')T(s,a,s')\Big]\|b_1 - b_2\|_\infty. \end{aligned}$$

D'où le résultat attendu. □

En pratique, on utilisera comme constante de Lipschitz $\mu = \max_{a,z} \mu_{a,z}$.

Des deux lemmes précédents, on tire le corollaire suivant (où la norme utilisée est toujours la norme infinie).

Corollaire 1
Sur un horizon fini H, la fonction de valeur optimale est ν_t-lipschitzienne pour tout $t \in \{0,\dots,H\}$, où

$$\nu_H = 0, \quad \text{et pour } t \in \{1,\dots,H\},$$
$$\nu_{t-1} = (\gamma\lambda)\nu_t + \left(\rho + \gamma|\mathcal{Z}|V^{lim}\mu\right).$$

Démonstration 8
Pour $t = H$, on a trivialement $|V_H^*(b_1) - V_H^*(b_2)| = 0$, donc la propriété est vérifiée.

Supposons maintenant que la propriété est vérifiée pour $t \in \{1,\dots,H\}$. On a alors, pour tout (b_1,b_2) :

$$V_{t-1}^*(b_1) = \max_a \Big[r(b_1,a) + \gamma \sum_z P(z|b_1,a)V_t^*(b_1^{a,z})\Big].$$

Or, V_t étant lipschitzienne, avec les lemmes 6 et 7, et en appliquant le lemme 2 :

$$P(z|b_1,a)\cdot V_t^*(b_1^{a,z}) \leq P(z|b_2,a)\cdot V_t^*(b_2^{a,z}) + \Big(\underbrace{|V_t^*(b_1^{a,z})|}_{\leq V^{lim}}\cdot\mu + |P(z|b_1,a)|\cdot\nu_t\lambda\Big)\|b_1 - b_2\|.$$

D'où,

$$\begin{aligned} &V_{t-1}^*(b_1) \\ &\leq \max_a \Big[r(b_2,a) + \rho\|b_1 - b_2\| + \gamma\sum_z \left(P(z|b_2,a)V_t^*(b_2^{a,z}) + \left(V_t^{lim}\mu + P(z|b_1,a)\nu_t\lambda\right)\|b_1 - b_2\|\right)\Big] \\ &\leq \max_a \Big[r(b_2,a) + \gamma\sum_z \left(P(z|b_2,a)V_t^*(b_2^{a,z})\right) \\ &\qquad + \rho\|b_1 - b_2\| + \gamma\sum_z \left(V_t^{lim}\mu\right)\|b_1 - b_2\| + \gamma\underbrace{\sum_z P(z|b_1,a)}_{=1}\nu_t\lambda\|b_1 - b_2\|\Big] \\ &\leq \max_a \Big[r(b_2,a) + \gamma\sum_z \left(P(z|b_2,a)V_t^*(b_2^{a,z})\right) + \left(\rho + \gamma|\mathcal{Z}|V_t^{lim}\mu + \gamma\nu_t\lambda\right)\|b_1 - b_2\|\Big] \\ &\leq V_{t-1}^*(b_2) + \underbrace{\left((\gamma\lambda)\nu_t + \left(\rho + \gamma|\mathcal{Z}|V_t^{lim}\mu\right)\right)}_{\nu_t}\|b_1 - b_2\|. \end{aligned}$$

On a donc bien la formule de récurrence attendue. □

TABLE 1 – Analogies entre MDP s-lipschitziens stochastiques et (ρ)-POMDP

MDP	(ρ-)POMDP
τ	λ
ϕ	μ
$vol(\mathcal{X})$	$\lvert\mathcal{Z}\rvert$

Comme dans le cas des MDP stochastiques, on peut dériver une suite de constantes de Lipschitz plus grossière mais dont on peut obtenir une expression non récursive.

Corollaire 2
Sur un horizon fini H, la fonction de valeur optimale vérifie, $\forall b_1, b_2, \forall t \in \{0, \dots, H\}$,

$$|V_t^*(b_1) - V_t^*(b_2)| \leq \nu_t^{max} \|b_1 - b_2\|,$$
$$\text{où } \nu_t^{max} = \begin{cases} \left(\rho + \gamma V^{lim} \mu |\mathcal{Z}|\right) \cdot \dfrac{1 - (\gamma\lambda)^{H-t}}{1 - (\gamma\lambda)} & \text{si } \gamma\lambda \neq 1; \\ \left(\rho + \gamma V^{lim} \mu |\mathcal{Z}|\right) \cdot (H - t) & \text{si } \gamma\lambda = 1. \end{cases}$$

Démonstration 9
Comme précédemment, la propriété est trivialement vérifiée pour $t = H$.

En remplaçant, pour tout $t \in \{1, \dots, H\}$, V_t^{lim} par $V^{lim} (> V_t^{lim})$ dans l'expression reliant ν_{t-1} à ν_t, on construit une suite $(\nu_t^{max})_{t \in \{0,\dots,H\}}$ vérifiant, pour tout $t \in \{1, \dots, H\}$,

$$\nu_{t-1}^{max} = \Big(\underbrace{(\gamma\lambda)}_{\alpha} \nu_t^{max} + \underbrace{\left(\rho + \gamma |\mathcal{Z}| V^{lim} \mu\right)}_{\beta} \Big) \|b_1 - b_2\|.$$

On peut alors appliquer le lemme 1 pour obtenir le résultat escompté pour tout t.

Comme on pouvait s'y attendre, on a les analogies présentées dans la table 1 avec le cas des MDP s-lipschitziens stochastiques.

5 Algorithmes

Nous allons maintenant discuter de différents algorithmes possibles selon que l'horizon est fini ou non, et qu'un état de croyance initial est fourni ou non. Les résultats sont présentés dans le cadre des MDP continus stochastiques, mais transposables directement aux (ρ)POMDP. On fait toutefois l'hypothèse que le nombre d'états atteignables suite à l'application d'une action est fini et borné, de manière à remplacer les intégrales par des sommes finies.

5.1 Préliminaire : Approximation d'une fonction lipschitzienne

Une étape préliminaire est l'approximation d'une simple fonction k-lipschitzienne $f : X \to \mathbb{R}$ à erreur bornée $\epsilon > 0$ ($k \in \mathbb{R}^+$). Soit X' un ensemble de N points de X. Si la valeur de f n'est connue qu'en les points de X', alors pour tout $x \in X$, on a :

$$\underbrace{\max_{x' \in X'} \left[f(x') - k \cdot d(x, x')\right]}_{L_f^{X'}(x)} \leq f(x) \leq \underbrace{\min_{x' \in X'} \left[f(x') + k \cdot d(x, x')\right]}_{U_f^{X'}(x)}.$$

Désormais, on considérera essentiellement les deux types d'approximateurs (minorant et majorant) que sont $L_f^{X'}$ et $U_f^{X'}$. La figure 1 ne correspond pas tout à fait à cette situation car d'une part les valeurs exactes de référence $f(x)$ y sont remplacées par des valeurs majorantes (pour $U_f^{X'}$) ou minorantes (pour $L_f^{X'}$), et d'autre part des hyperplans constants sont aussi employés.

A l'évidence, pour que l'incertitude sur la valeur de f soit bornée, il faut que la distance maximale entre tout point de X et l'ensemble X' soit majorée : $\max_{x \in X} \min_{x' \in X'} d(x, x') \leq \delta (\in \mathbb{R}^+)$. En le point x

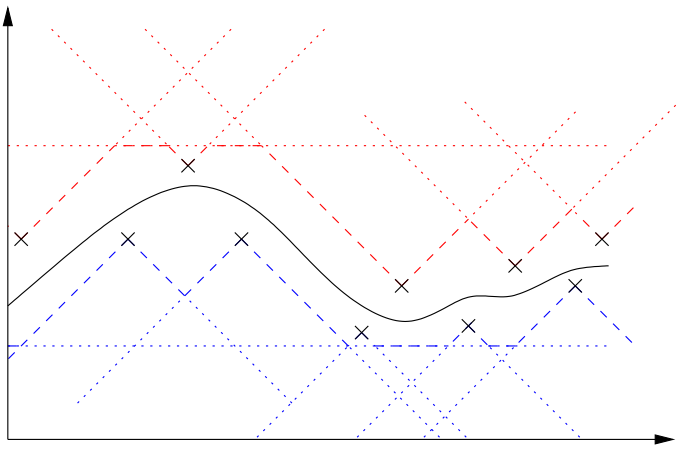

FIGURE 1 – Une fonction de valeur (mono-dimensionnelle) hypothétique et ses approximations en dents de scie (plus un hyperplan constant) majorantes (en rouge) et minorante (en bleu)

maximisant cette distance, l'incertitude $U_f^{X'}(x) - L_f X'(x)$ est majorée par $2k\delta$. L'existence de δ implique que le domaine X lui-même soit borné ($\max_{x,x'} d(x,x') \leq M (\in \mathbb{R}^+)$) pour qu'évaluer la fonction en un nombre fini de points suffise à obtenir une approximation bornée.

Etant donné X et une erreur souhaitée ϵ, le nombre minimum de points à sélectionner dans X' est le nombre de couverture de X pour ϵ, c'est à dire le nombre minimum de boules de rayons $\frac{\epsilon}{2k}$ (et ici de centres dans X) nécessaire pour couvrir X. La condition d'appartenance des centres à X pouvant compliquer la construction d'une telle couverture minimale, nous supposerons X convexe, hypothèse toujours vérifiée dans les bMDP. On peut espérer générer moins de points en construisant un ensemble X' de manière incrémentale, comme suit (on notera $X'_i = \{x_1, \ldots, x_i\}$),

— en prenant un premier point x_1 au hasard (ou minimisant $\max_{x \in X} d(x_1, x)$),

— puis en ajoutant tout nouvel x_{i+1} de manière à minimiser $U_f^{X'_i}(x_{i+1}) - L_f^{X'_i}(x_{i+1})$.

Cette méthode requiert toutefois de résoudre de multiples problèmes d'optimisation sur des ensembles continus, ce qui peut s'avérer coûteux.

Une approche inspirée de Munos (2011, 2014) consisterait à utiliser un partitionnement hiérarchique ou, en d'autres termes, un partitionnement de plus en plus fin de X, partitionnement dans lequel chaque sous-ensemble est associé à de simples majorants et minorant de f déductibles de la taille de ce sous-ensemble et de la constante de Lipschitz. En exploitant la capacité à calculer des constantes de Lipschitz locales, on peut obtenir un partitionnement plus ou moins fin selon la région de X. Par contre on n'échappe pas à la nécessité d'avoir une approximation aussi bonne sur tout X, alors que Munos (2011) n'a besoin que de garantir qu'une solution ϵ-optimale de f est trouvée.

5.2 Opérateur de Bellman

L'opérateur (d'optimalité) de Bellman, noté ici $\mathcal{H}$, met normalement à jour la fonction de valeur pour tout état (que l'horizon soit fini ou infini), ce qui ne sera pas possible en pratique dans nos cadres continus. Les familles d'approximateurs employées ici (minorant ou majorant) vont toutefois nous permettre à la place de mettre à jour la fonction de valeur en un ensemble de points à la fois (ceux de $\mathcal{S}'$) – opérateur $\mathcal{H}^{\mathcal{S}'}$ – ou en un seul point s – opérateur $\mathcal{H}^s$, mises à jour qui auront un impact sur toute une région de $\mathcal{S}$.

5.3 Programmation dynamique

Supposons d'abord qu'aucun état (de croyance) initial n'est défini, et que l'horizon temporel est fini.

La section précédente nous fournit deux types d'approximateurs à base d'un ensemble de points d'évaluation, l'un minorant, l'autre majorant.

— Si on choisit d'être pessimiste, on utilisera des approximateurs minorants L_t que l'on complétera par le minorant toujours valable V_t^{min}. Si, pour t donné, on note $\mathcal{S}'_t$ l'ensemble des points utilisés dans L_t, et $l_t(s)$ la valeur associée à un de ces points, alors $L_t(s) \stackrel{\text{def}}{=} \max\left\{V_t^{min}, \max_{s' \in S_t} \left(l_t(s') - \nu_t d(s,s')\right)\right\}$.

— Si on choisit d'être optimiste, on utilisera des approximateurs majorants U_t que l'on complétera par le majorant toujours valable V_t^{max}. Si, pour t donné, on note $\mathcal{S}'_t$ l'ensemble des points utilisés dans L_t, et $u_t(s)$ la valeur associée à un de ces points, alors $U_t(s) \stackrel{\text{def}}{=} \min\{V_t^{max}, \min_{s' \in S_t}(u_t(s') + \nu_t d(s,s'))\}$.

La figure 1 illustre la fonction de valeur optimale et ces deux fonctions à une étape t, en illustrant le fait que les valeurs de référence $l_t(s)$ (respectivement $u_t(s)$) pour $s \in \mathcal{S}'_t$ ne sont pas les valeurs exactes $V_t^*(s)$, mais des minorants (resp. des majorants). Par la suite, sauf besoin particulier, on notera simplement l'approximateur utilisé V_t (qu'on suive une approche optimiste ou pessimiste), et la valeur en un point s de $\mathcal{S}'_t$ sera $v_t(s)$.

Pour $t = H$, on a trivialement $\mathcal{S}'_H = \emptyset$. Comme $V_H^{min} = V_H^{max} = 0$, l'approximateur V_H est toujours exact. Pour t de $H-1$ à 0, on construit les approximateurs V_t successivement en calculant la valeur $v_t(s)$ de chaque point de l'ensemble $\mathcal{S}'_t$ par application de l'opérateur de Bellman $\mathcal{H}^s$ sur l'approximateur V_{t+1} défini sur l'ensemble de l'espace d'état :

$$v_t(s) \stackrel{\text{def}}{=} \max_a \Big(r(s,a) + \gamma \sum_{s'} T(s,a,s') V_{t+1}(s') \Big).$$

Cette approche est similaire à celle employée pour les POMDP pour encadrer la fonction de valeur PWLC à t par une enveloppe d'hyperplans (comme minorant) et un approximation en dents de scies (comme majorant). Comme pour les POMDP, éliminer régulièrement les éléments inutiles des approximateurs permettra de gagner du temps de calcul (et de la mémoire). Ici, on éliminera des points des ensembles $S_{U,t}$ et $S_{L,t}$ s'ils sont *inutiles*, c'est-à-dire :

$$(s, u_t(s)) \text{ est inutile si } \begin{cases} u_t(s) \geq V_t^{max} \text{ ou} \\ \exists s' \in S_{U,t} \setminus \{s\} \text{ tel que } u_t(s) \geq u_t(s') + \nu_t d(s,s'); \end{cases}$$

$$(s, l_t(s)) \text{ est inutile si } \begin{cases} l_t(s) \leq V_t^{min} \text{ ou} \\ \exists s' \in S_{L,t} \setminus \{s\} \text{ tel que } l_t(s) \leq l_t(s') - \nu_t d(s,s'). \end{cases}$$

Le choix des points à mettre dans $\mathcal{S}'_t$ reste une difficulté. Comparé à l'approximation d'une fonction simple, on va ici cumuler des erreurs d'approximation. Supposons que, quelle que soit la méthode de sélection des ensembles $\mathcal{S}'_t$ choisie, elle garantisse que l'erreur de l'approximateur est au plus de $\epsilon' > 0$ (indépendamment des erreurs cumulées aux itérations précédentes). On a alors le résultat suivant.

Théorème 1 (∼Théorême 3.1 de Pineau *et al.* (2006))
Pour tout t, l'erreur de la programmation dynamique $\epsilon_t = \|V_t - V_t^*\|_\infty$ est majorée par

$$\epsilon_t \leq \frac{1-\gamma^{H-t}}{1-\gamma}\epsilon'.$$

Démonstration 10

$$\begin{aligned}
\epsilon_t &= \|V_t - V_t^*\|_\infty \\
&= \|\mathcal{H}^{\mathcal{S}'_t} V_{t+1} - \mathcal{H} V_{t+1}^*\|_\infty && \textit{(par définition de } \mathcal{H}^{\mathcal{S}'_t}\textit{)} \\
&\leq \|\mathcal{H}^{\mathcal{S}'_t} V_{t+1} - \mathcal{H} V_{t+1}\|_\infty + \|\mathcal{H} V_{t+1} - \mathcal{H} V_{t+1}^*\|_\infty && \textit{(par inégalité triangulaire)} \\
&\leq \epsilon' + \|\mathcal{H} V_{t+1} - \mathcal{H} V_{t+1}^*\|_\infty && \textit{(par hypothèse)} \\
&\leq \epsilon' + \gamma \|V_{t+1} - V_{t+1}^*\|_\infty && \textit{(par contraction de la mise à jour exacte)} \\
&= \epsilon' + \gamma \epsilon_{t+1} && \textit{(par définition de } \epsilon_{t+1}\textit{)} \\
&\leq \frac{1-\gamma^{H-t}}{1-\gamma}\epsilon'. && \textit{(par sommation d'une série géométrique)}
\end{aligned}$$

□

En notant ν le maximum de ν_t sur $t \in \{0, \ldots, H\}$ et en prenant par exemple des ensembles $\mathcal{S}'_t$ formant une δ-couverture de $\mathcal{S}$, on a $\epsilon' \leq \delta\nu$ et $\epsilon_h \leq \frac{1-\gamma^{H-t}}{1-\gamma}\delta\nu$.

Pour obtenir une erreur ϵ_0 donnée, ce qui est souvent le critère premier, on peut donc employer $\epsilon' = \left(\frac{1-\gamma^H}{1-\gamma}\right)^{-1}\epsilon_0$ à chaque étape (pour chaque $t \in \{0, \ldots, H-1\}$). Si on construit à l'avance les ensembles de point $\mathcal{S}'_t$ (et non incrémentalement), alors le même ensemble $\mathcal{S}'$ peut être repris pour tout t.

Approximation d'un horizon temporel infini

Dans le cas d'un problème à horizon temporel infini, en l'absence de garantie que la suite de constantes de Lipschitz (ν_t) converge, il est préférable de se ramener à un problème à horizon fini. Pour garantir une erreur inférieure à $\epsilon > 0$, on peut choisir ϵ' et H par exemple de sorte que

$$\epsilon' = \left(\frac{1-\gamma^H}{1-\gamma}\right)^{-1} \frac{\epsilon}{2}$$

$$\text{et} \qquad \gamma^H \frac{R_{max}}{1-\gamma} \leq \frac{\epsilon}{2} \qquad \text{(pour que ce qui se passe au-delà de } H \text{ soit négligeable)}$$

$$\text{c'est-à-dire} \qquad \gamma^H \leq \frac{\epsilon}{2}\frac{1-\gamma}{R_{max}}$$

$$exp\left(H \ln \gamma\right) \leq exp\left(\ln\left(\frac{\epsilon}{2}\frac{1-\gamma}{R_{max}}\right)\right)$$

$$H = \left\lceil \frac{\ln\left(\frac{\epsilon}{2}\frac{1-\gamma}{R_{max}}\right)}{\ln \gamma} \right\rceil .$$

5.4 Itération sur la valeur

Approcher un problème à horizon infini en tronquant l'horizon peut être très coûteux en mémoire si l'horizon employé est grand. Si la suite des constantes de Lipschitz (ν_t) converge, on peut alors préférer employer un algorithme d'itération sur la valeur, par exemple synchrone. On indicera alors les fonctions de valeur par h pour désigner l'*horizon* qui a été considéré jusque là.

L'algorithme d'itération sur la valeur revient alors, en ayant par exemple choisi une δ-couverture $\mathcal{S}'$ de $\mathcal{S}$, à initialiser V_0, puis à calculer itérativement $V_h = \mathcal{H}^{cS'} V_{h-1}$ pour tout $h \geq 1$. Les bornes d'erreurs vues précédemment pour la programmation dynamique s'étendent de manière immédiate.

5.5 HSVI

Les algorithmes présentés jusqu'ici ont deux défauts importants :

— ils requièrent de construire des ensembles de points $\mathcal{S}'$ pour couvrir suffisamment bien l'espace d'états, ce qui implique soit une répartition régulière d'un grand nombre de points, soit une construction incrémentale (avec des optimisations continues à chaque étape) d'un nombre un peu moins important de points ; et

— ils n'exploitent pas la connaissance possible d'un état (de croyance) initial, ce qui permettrait de concentrer les efforts de calcul sur les parties atteignables de l'espace d'état.

Nous allons ici proposer une adaptation de l'algorithme *Heuristic Search Value Iteration* (HSVI) de Smith & Simmons (2004, 2005); Smith (2007), laquelle va permettre de pallier (pour partie) ces défauts. Evidemment, d'autres algorithmes à base de points (PBVI (Pineau *et al.*, 2003), PERSEUS (Spaan & Vlassis, 2005), FSVI (Shani *et al.*, 2007), SARSOP (Kurniawati *et al.*, 2008), GapMin (Poupart *et al.*, 2011) ...) pourraient être envisagés. Nous optons pour HSVI entre autres parce qu'il est relativement simple et permet de contrôler l'erreur faite (du fait de l'utilisation conjointe d'un minorant et d'un majorant).

Comme dans la version originale, nous considérons ici le cas d'un horizon temporel infini [5]. Le problème est caractérisé par un état initial s_0 et une erreur maximum souhaitée en s_0 bornée par $\epsilon > 0$.

5.5.1 Approximateurs uniformément améliorables

On remarquera d'abord que, avec une initialisation minorante (pour L_0) ou majorante (pour U_0), les approximateurs employés ici sont en outre *uniformément améliorables* (Zhang & Zhang, 2001; Smith, 2007), c'est-à-dire que, pour tout h :

$$L_h \leq \mathcal{H}L_h \leq V^* \leq \mathcal{H}U_h \leq U_h,$$

5. Il est toutefois possible de déterminer une politique à horizon temporel fini comme cela a été fait pour des DecPOMDP (Dibangoye *et al.*, 2013, 2016).

ce qui implique aussi :

$$L_h \leq L_{h+1} \leq V^* \leq U_{h+1} \leq U_h.$$

Ainsi, chaque itération ne peut faire qu'améliorer l'approximation de V^*.

Nous partons des travaux de Hauskrecht (1997, 2000); Smith (2007) pour chercher des initialisations possibles L_0 et U_0 plus fines que V^{min} et V^{max} dans le cas de ρ-POMDP non convexes. Pour L_0, faire une recherche directe de politique dans un espace de politiques restreint (par exemple celui des politiques aveugles ou à action fixe) et évaluer la politique trouvée est une méthode applicable. L'évaluation de la politique sera toutefois plus coûteuse que dans le cas d'un POMDP (avec les enveloppes d'alpha-vecteurs). Pour U_0, une approche courante est de mettre en œuvre un algorithme d'itération sur la valeur avec un opérateur "optimiste" par rapport à l'opérateur de Bellman. Toutefois, les méthodes "MDP", "QMDP" et "FIB" (*fast informed bound*) ne sont plus valables, du fait de la non-convexité de la fonction de valeur. Un moyen de pouvoir quand même s'y ramener serait de majorer la fonction de récompense par une fonction linéaire et de calculer un majorant de la fonction de valeur du POMDP ainsi obtenu.

5.5.2 L'algorithme s-Lipschitz HSVI

Les approximateurs employés U et L étant des minorants et majorants uniformément améliorables, on peut reprendre le schéma d'HSVI (voir algorithme 1), lequel génère des trajectoires

— en partant de l'état initial s_0 ;
— en choisissant, dans chaque état s, l'action a maximisant $Q^U(s,a) = r(s,a) + \gamma \sum_{s'} T(s,a,s')U(s')$;
— en choisissant le prochain état s' maximisant $T(s,a,s')\,(U(s') - L(s'))$;
— en mettant à jour la fonction de valeur localement en chaque s visité ; et
— en s'arrêtant quand $\gamma^d(U(s) - L(s)) \leq \epsilon$, où d est la profondeur actuelle de la trajectoire en cours.

Algorithme 1 : s-Lipschitz Heuristic Search Value Iteration (sL-HSVI)

Fct `HSVI` (ϵ)
  Initialiser L et U (par exemple avec V^{max} et V^{min})
  Calculer ν_∞^{max}
  si $\nu_\infty^{max} = \infty$ **alors**
    retourner *Erreur*
  tant que $(U(s_0) - L(s_0)) > \epsilon$ **faire**
    `EssayerRécursivement` $(s_0, d = 0)$
  retourner L
Fct `EssayerRécursivement` (s, d)
  si $\gamma^d(U(s) - L(s)) > \epsilon$ **alors**
    `MettreAJour` (s)
    $a^* \in \arg\max_{a \in \mathcal{A}} Q^U(s,a)$
    $s^* \in \arg\max_{s' \in Suivants(s,a^*)} Pr(s,a^*,s')(U(s') - L(s'))$
    `EssayerRécursivement` $(s^*, d+1)$
    `MettreAJour` (s)
  retourner
Fct `MettreAJour` (s)
  $L \leftarrow$ miseAJour(L, s)
  $U \leftarrow$ miseAJour(U, s)

Une des spécificités de l'algorithme *s-Lipschitz HSVI* (sL-HSVI) obtenu est de d'abord calculer la constante de Lipschitz ν_∞^{max}, et de s'interrompre si cette limite diverge. Une autre spécificité est évidemment l'utilisation des fonctions majorantes et minorantes en dents de scie lipschitziennes de constante ν_∞^{max}. La fonction `MettreAJour`(s) effectue une mise à jour locale de U comme de L. Elle calcule $u(s) = \max_a r(s,a) + \gamma \sum_{s'} T(s,a,s')U(s')$ (et, en même temps, l'action la plus prometteuse a^*) et $l(s) = \max_a r(s,a) + \gamma \sum_{s'} T(s,a,s')L(s')$ pour ajouter les points $(s, u(s))$ et $(s, l(s))$ aux deux approximateurs (sauf s'ils atteignent V^{max} ou V^{min}).

On notera par ailleurs :
— qu'une procédure d'élagage de l'ensemble S_U (resp. de S_L) peut être appliquée quand la taille de cet ensemble dépasse un certain seuil (par exemple lorsque cet ensemble a cru de 10% depuis le dernier élagage) ; et
— que, dans le cas de problèmes à horizon temporel fini, on calculera une fonction de valeur par pas de temps, ce qui passera par la détermination d'une constante ν_t pour chaque pas de temps par récurrence (ce qui permettra des approximations plus fines) comme vu dans le lemme 4 et le corollaire 1.

Une question intéressante est de savoir s'il est préférable de traiter les problèmes à horizon infini comme présenté ci-dessus, ou en les approchants avec un horizon temporel fini, ce qui permet d'avoir des approximateurs plus efficaces (du fait des constantes de Lipschitz plus fines).

Le schéma algorithmique employé ici reste identique à celui introduit par Smith & Simmons (2004). Les propriétés théoriques l'accompagnant sont par ailleurs préservée (du fait de l'uniforme améliorabilité de U et L). Ainsi l'algorithme converge en un nombre fini d'itérations.

6 Discussion et conclusion

Nous avons ici montré que, moyennant l'hypothèse que son modèle (T et r) était lipschitzien par rapport à l'état, on pouvait approcher la fonction de valeur d'un MDP à espace d'état continu par une fonction lipschitzienne en dents de scie tout en bornant l'erreur faite. A l'instar de la propriété de convexité et de linéarité par morceaux dans les POMDP, cette propriété permet de proposer des algorithmes de résolution à erreur bornée pour ces MDP lipschitziens. Ceci s'applique en particuliers aux ρ-POMDP – variantes de POMDP pour la recherche d'information –, problèmes dans lesquels la fonction de valeur n'est pas nécessairement convexe dans l'espace des états de croyance. Ces approximations par des fonctions en dent de scie lipschitziennes sont nettement moins fines que les approximations (pseudo-)convexes [6] employées pour les POMDP, mais des expérimentations seraient nécessaires pour mieux évaluer la dégradation obtenue, en particulier en fonction de la dimensionalité du problème. Des expérimentations permettraient par ailleurs de juger de la finesse des majorants des constantes de Lipschitz obtenus.

Une perspective intéressante serait aussi d'approcher (à erreur bornée) un modèle non-lipschitzien de MDP par un modèle lipschitzien. On pourrait alors appliquer des algorithmes de résolution tels que présentés ici tout en majorant l'erreur faite en les employant sur un modèle approché plutôt que sur le problème original. Une autre perspective serait de lever l'hypothèse du facteur de branchement fini (utile pour les MDP), par exemple en employant des algorithmes reposant sur de l'échantillonnage de trajectoires, donc des garanties moins fortes. Une inspiration pourrait être *Forward Search Sparse Sampling* (FSSS) de Walsh *et al.* (2010), un algorithme cousin de Sparse Sampling (SS) (Kearns *et al.*, 2002) et de MCTS (Coulom, 2006; Chaslot *et al.*, 2008), mais employant des estimations majorantes et minorantes de la fonction de valeur comme HSVI. En revanche FSSS n'exploite pas de généralisation de ces minorants et majorants comme le fait HSVI. Enfin, on notera que la Lipschitz-continuité par rapport à un espace d'actions continu est aussi une propriété que l'on pourrait aborder (comme l'ont fait Ieong *et al.* (2007)), par exemple en utilisant des algorithmes d'optimisation spécifiques tels que proposés par Munos (2011, 2014), et toujours en majorant l'erreur faite.

Mais ces travaux ont d'abord été motivés par le souhait de résoudre des jeux de Markov à observabilité partielle (POSG). En effet, si la transformation en un *occupancy MDP* d'un DecPOMDP a permis à Dibangoye *et al.* (2013, 2016) de se ramener à un cadre comparable à celui des bMDP (avec fonction de valeur PWLC), la même transformation, appliquée à un POSG, devrait permettre de se ramener à un jeu de Markov (complètement observable) pour lequel la fonction de valeur (scalaire dans un jeu à 2 joueurs et somme nulle, vectorielle sinon) devrait être non-convexe, mais lipschitzienne.

Références

ARAYA-LÓPEZ M., BUFFET O., THOMAS V. & CHARPILLET F. (2010). A POMDP extension with belief-dependent rewards. In *Advances in Neural Information Processing Systems 23 (NIPS-10)*.

ASTROM K. (1965). Optimal control of Markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, **10**(1), 174 – 205.

6. *Saw-tooth* n'est pas convexe.

BELLMAN R. (1957). A Markovian decision process. *Journal of Mathematics and Mechanics*, **6**(5), 679–684.

CHASLOT G., BAKKES S., SZITA I. & SPRONCK P. (2008). Monte-Carlo tree search : A new framework for game AI. In *Proceedings of the Fourth Artificial Intelligence and Interactive Digital Entertainment Conference (AIIDE)*.

COULOM R. (2006). Efficient selectivity and backup operators in Monte-Carlo tree search. In *Proceedings of the Fifth International Conference on Computer and Games (CG-2006)*.

DIBANGOYE J., AMATO C., BUFFET O. & CHARPILLET F. (2013). Optimally solving Dec-POMDPs as continuous-state MDPs. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence (IJCAI-13)*.

DIBANGOYE J., AMATO C., BUFFET O. & CHARPILLET F. (2016). Optimally solving Dec-POMDPs as continuous-state MDPs. *Journal of Artificial Intelligence Research*, **55**, 443–497.

HAUSKRECHT M. (1997). Incremental methods for computing bounds in partially observable Markov decision processes. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI)*.

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

IEONG S., LAMBERT N., SHOHAM Y. & BRAFMAN R. (2007). Near-optimal search in continuous domains.

KEARNS M., MANSOUR Y. & NG A. (2002). A sparse sampling algorithm for near-optimal planning in large Markov decision processes. *Machine Learning*, **49**, 193–208.

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*.

MUNOS R. (2011). Optimistic optimization of a deterministic function without the knowledge of its smoothness. In *Advances in Neural Information Processing Systems (NIPS'11)*.

MUNOS R. (2014). From bandits to Monte-Carlo Tree Search : The optimistic principle applied to optimization and planning. *Foundations and Trends in Machine Learning*, **7**(1), 1–130.

PINEAU J., GORDON G. & THRUN S. (2003). Point-based value iteration : An anytime algorithm for POMDPs. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI)*, p. 1025–1032.

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

PLATZMAN L. K. (1977). *Finite Memory Estimation and Control of Finite Probabilistic Systems*. PhD thesis, Massachusetts Institute of Technology (MIT).

POUPART P., KIM K.-E. & KIM D. (2011). Closing the gap : Improved bounds on optimal POMDP solutions. In *Proceedings of the Twenty-First International Conference on Automated Planning and Scheduling (ICAPS)*.

SHANI G., BRAFMAN R. & SHIMONY S. (2007). Forward search value iteration for POMDPs. In *Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence (IJCAI'07)*.

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. (2007). *Probabilistic Planning for Robotic Exploration*. PhD thesis, The Robotics Institute, Carnegie Mellon University.

SMITH T. & SIMMONS R. (2004). Heuristic search value iteration for POMDPs. In *Proceedings of the Annual Conference on Uncertainty in Artificial Intelligence (UAI)*.

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

SOMANI A., YE N., HSU D. & LEE W. S. (2013). DESPOT : Online POMDP planning with regularization. In *Advances in Neural Information Processing Systems (NIPS)*.

SONDIK E. (1971). *The Optimal Control of Partially Observable Markov Decision Processes*. PhD thesis, Standford University.

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

WALSH T., GOSCHIN S. & LITTMAN M. (2010). Integrating sample-based planning and model-based reinforcement learning. In *Proceedings of the National Conference on Artificial Intelligence (AAAI'10)*.

ZHANG N. L. & ZHANG W. (2001). Speeding up the convergence of value iteration in partially observable Markov decision processes. *Journal of Artificial Intelligence Research*, **14**, 29–51.