Une approche du traitement du temps dans le cadre MDP : trois méthodes de découpage de la droite temporelle

E. RACHELSON(*) P. FABIANI(*) J.-L. FARGES(*) F. TEICHTEIL(*)
F. GARCIA(**)

(*) ONERA-DCSD ; 2, Avenue E. Belin BP 4025 31055 Toulouse cedex 4
{*emmanuel.rachelson, patrick.fabiani, jean-loup.farges, florent.teichteil*}*@cert.fr*
(**) INRA-BIA Chemin de Borde-Rouge-Auzeville BP 52627 F-31326 Castanet-Tolosan cedex
fgarcia@toulouse.inra.fr

Résumé

De nombreux problèmes de planification s'inscrivent dans un environnement instationnaire. Dans le cadre de la décision dans l'incertain sur horizon infini, pour les problèmes stationnaires à l'infini, on se propose de définir un cadre de modélisation dérivé du modèle SMDP dans lequel la variable temporelle est observable par l'agent. Dans ce cadre, nous développons trois approches de résolution différentes afin de générer des politiques qui en tout état discret du système spécifient l'action optimale à entreprendre en fonction de la date courante.

1 Introduction

Les problèmes de planification, dans un cadre déterministe ou incertain, sont souvent considérés comme des problèmes stationnaires, où les notions de changement d'état, de récompense, ou de coût n'ont pas de dépendance directe à la variable temporelle. On s'intéresse ici au traitement des problèmes dépendant explicitement du temps. Cette approche est motivée par un problème plus complexe dans le cadre de la coopération bi-agents mais les résultats s'appliquent dans le cadre général de la construction d'un plan dépendant explicitement du temps. Les problèmes de coopération et de coordination de deux agents indépendants impliquent un échange d'information concernant l'impact de l'action prévue d'un des agents sur l'univers de l'autre ([CC03], [TC04], [Bou96], [BM05], [Rac05]). On se place dans un cadre où deux agents échangent une représentation compacte de leur stratégie pour pouvoir planifier chacun de leur côté. Le problème, par agent, se ramène à un problème mono-agent de décision dans l'incertain présentant des dépendances temporelles dues aux déclarations d'intentions de l'autre agent. Dans ce cadre, on cherche un moyen de construire des plans robustes dépendant du temps. Cela implique que la variable temporelle soit observable du point de vue de l'agent, *ie.* qu'elle appartienne à son espace d'état. De façon plus générale, on aboutit au problème de la planification mono-agent présentant un certain nombre d'échéances et dont certains paramètres varient avec la variable temporelle.

Dans un premier temps (section 1) on cherchera à cerner et circonscrire le problème sans se soucier de modélisation, puis on introduira un cadre formel basé sur le modèle SMDP afin de pouvoir chercher des solutions sous forme de politiques (section 2). On présentera alors trois approches permettant d'obtenir une politique de la forme $\pi(s,t)$ dont on discutera les avantages et les faiblesses (sections 3, 4 et 5). On concluera enfin sur l'implémentation en cours et les intérêts comparés des trois méthodes (section 6).

1.1 Le type de problèmes traités

On va chercher ici à définir plus précisément les problèmes auxquels on s'intéresse. On cherche à travailler sur un problème dont le modèle d'univers peut dépendre du temps. Par modèle d'univers, on entend l'ensemble des données connues de l'agent sur lesquelles il s'appuie pour construire son plan. Dans ce cadre, on cherche des plans robustes aux incertitudes d'action. Les échéances du problème se propagent dans le modèle d'univers et font qu'une décision qui est optimale à un instant donné ne l'est plus à un autre, on est mené pour celà à définir des *dates de décision*. Une date de décision est une date - observable par l'agent - en laquelle l'agent a défini une action de son plan.
Par ailleurs, on se place dans un cadre de décision dans l'incertain : les résultats des actions de l'agent sur l'univers ne sont pas déterministes, elles peuvent avoir plusieurs issues, que l'on peut prendre en compte dans le modèle d'univers de l'agent.
Enfin, on s'intéresse à des problèmes dont l'évolution prévue dans le temps est limitée dans l'avenir. On nomme ces problèmes "stationnaires à l'infini" car on peut définir une date à partir de laquelle la dépendance au temps s'arrête et le problème devient stationnaire. On appelle cette date le *pseudo-horizon*.

1.2 Un exemple représentatif

Les problèmes correspondant à ce type de modélisation sont des problèmes présentant des aspects de coordination. Un exemple représentatif de ces problèmes est une extension du problème du taxi ([Die98]) où deux taxis sont présents sur un monde grille où il y a un passager, une pompe à essence, et un point de dépot. Du point de vue de chaque agent, ce problème revient à résoudre un problème mono-agent où le passager disparaît à une date t_d, date à laquelle l'autre agent l'a pris (la récompense est alors automatiquement acquise). Du point de vue de l'opérateur humain, une bonne stratégie consiste à dire "jusqu'à une date peu avant t_d, aller chercher le passager", "ne rien faire au-delà de cette date". On cherche à obtenir des politiques, solutions du problème, aussi compactes que ces assertions et dépendant du temps en chaque état.

2 Formalisation du problème de la décision dans l'incertain en fonction du temps

Dans cette section on cherche à se munir d'un cadre formel pour traiter les problèmes dépendant du temps. On part du modèle des Processus Décisionnels de Markov (MDP) ([Put94]), modèle utile pour générer des plans robustes aux incertitudes d'action (2.1). On enrichit ce modèle d'une variable de durée selon le modèle SMDP présenté dans [Put94] (2.2). On adapte enfin ce modèle pour réécrire l'équation de Bellman [Bel57] avec une variable temporelle observable et continue (2.3). Dans ce cadre, on écrit au final le problème que l'on souhaite résoudre (2.4).

2.1 Le cadre MDP

Le modèle MDP est particulièrement adapté aux problèmes de décision dans l'incertain. Un problème défini sous forme MDP comporte :

- Un espace d'état S,
- Un jeu d'actions A,
- Un modèle de transition markovien $P(s'|s,a)$,
- Un modèle de récompense $r(s,a)$.

Sur ce modèle, on définit des plans sous formes de politiques. Une politique est une application de S dans A qui à chaque état du système associe une action à entreprendre. Afin de caractériser les politiques, on se dote de critères d'optimalité permettant de définir la fonction de valeur $V^\pi(s)$ d'une politique $\pi(s)$. Pour planifier sur un horizon infini on utilise le critère γ-pondéré ([Put94]) et on définit l'opérateur L_π :

$$L_\pi V(s) = r(\pi(s), s) + \gamma \sum_{s' \in S} P(s'|s, \pi(s)) V(s') \quad (1)$$

On peut alors montrer que V^π est l'unique point fixe de L_π. De même, on définit l'opérateur de programmation dynamique L :

$$LV(s) = \max_{a \in A} \left(r(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V(s') \right) \quad (2)$$

On peut alors définir l'équation de Bellman [Bel57] qui nous indique que la recherche de la politique optimale passe par la résolution du point fixe de l'opérateur L. On peut alors définir une mesure de l'écart à l'optimum sur la base de l'équation de Bellman : l'erreur de Bellman en s pour la politique π s'écrit :

$$BE(V^\pi(s)) = \max_{a \in A} \left(r(s,a) + \gamma \sum_{s' \in S} P(s,a,s') V^\pi(s') \right) - V^\pi(s) \quad (3)$$

Etant donné cette formulation récursive de la fonction de valeur optimale, on peut construire des algorithmes itératifs de recherche de la politique optimale. A partir du cadre MDP, on va essayer de prendre en compte l'impact de l'introduction d'une variable temporelle dans le modèle.

2.2 Le cadre SMDP

Dans un MDP "classique", chaque transition est implicitement considérée comme ayant une durée unitaire[1], cependant, les phénomènes réels sont de durée variable et les coûts associés sont souvent proportionnels à cette durée. On cherche donc à définir une distribution sur la durée des transitions et à l'intégrer au modèle. On effectue cette opération en enrichissant le modèle MDP de base. Le modèle des "Semi-Markov Decision Process" comporte :

- Un espace d'état S
- Un espace d'action discret A
- Un modèle de transition $Q(t, j|s, a)$ décrivant la probabilité que la prochaine date de décision se situe à t unités de temps dans l'avenir, et que le système soit dans l'état j sachant qu'on vient de prendre la décision de l'action a dans l'état s.
- Un modèle de récompense comportant des coûts ou gains immédiats à chaque transition notés $k(s,a)$ et des taux de récompense (ou de coût) $c(j,s,a)$.

Le modèle de transition est spécifié sous forme de deux fonctions P et F donnant respectivement la distribution sur l'état d'arrivée et la durée de transition de la transition (s,a). On a alors $Q(t, j|s, a) =$

[1] Cette durée unitaire est illustrée par le fait que dans l'équation de Bellman pour le critère γ-pondéré, on considère implicitement tous les instants de décision comme synchrones et que l'on multiplie les gains espérés à partir de l'étape $n+1$ par γ (et non γ^t), affirmant ainsi implicitement que l'on suppose que l'étape $n+1$ se situe à une unité de temps dans le futur.

$P(j|s,a) \cdot F(t|s,a)$ (indépendance de t et j).
Le modèle SMDP considère en fait deux processus distincts, le processus réel qui comporte tous les états de transition du système et le processus qu'on étudie (le SMDP lui-même). Ces deux processus concordent aux dates de décision. Le lien entre les deux processus est réalisé par la fonction $p(j|t,s,a)$ qui donne la probabilité que le processus réel soit dans l'état j, t unités de temps après avoir pris la décision a au point s. L'étude du modèle de récompense dans le cadre du processus réel permet de définir une fonction de récompense $r(s,a)$ sous la forme :

$$r(s,a) = k(s,a) + \int_0^\infty \sum_{j \in S} \left[\int_0^u \gamma^t c(j,s,a) p(j|t,s,a) dt \right] F(du|s,a) \quad (4)$$

Dans ce cadre, en se plaçant sur un horizon infini, en notant σ_n le temps écoulé en n étapes après le début de l'exécution, S_n la variable aléatoire décrivant l'état s, n étapes après le début de l'exécution et en se dotant d'un critère γ-pondéré, la fonction de valeur d'une politique π s'écrit : $V^\pi(s) = E_s^\pi \left(\sum_{n=0}^{\infty} \gamma^{\sigma_n} r_\pi(S_n) \right)$.
On peut montrer que cette expression se ramène à une formulation équivalente :

$$V^\pi(s) = r_\pi(s) + E_s^\pi \left(\gamma^{\sigma_1} V^\pi(S_1) \right) \quad (5)$$

Ce qu'on peut réécrire comme :

$$V^\pi(s) = r_\pi(s) + \sum_{j \in S} \int_0^\infty \gamma^t \cdot V^\pi(j) \cdot Q_\pi(dt, j|s) \quad (6)$$

Avec M_π la matrice $|S| \times |S|$ définie par : $m_\pi(j|s) = \int_0^\infty \gamma^t \cdot Q_\pi(dt, j|s)$, on peut se ramener ainsi dans un cadre similaire aux MDP classiques, écrire la valeur de la politique optimale comme unique solution de $V = \max_{\pi \in \mathcal{D}} \{r_\pi + M_\pi V\}$ et appliquer directement les algorithmes itératifs propres aux MDP classiques.

Ce modèle permet donc d'intégrer un coût temporel et des durées d'actions dans la résolution du MDP. Cependant les durées ne permettent pas de définir d'échéances si on ne se dote pas d'un temps avec une origine. En d'autres termes, on a introduit le temps dans le modèle MDP mais on ne l'a pas rendu observable, il faut donc enrichir encore le modèle et créer un modèle avec des dates obervables plutôt que des durées.

2.3 Un modèle avec dates

On enrichit donc notre modèle avec un axe du temps continu et une variable d'état supplémentaire t, qui vaut zéro à l'instant où l'agent effectue la planification et qui correspond, pour lui, à l'instant de début d'exécution. L'intérêt d'ajouter cette variable continue dans l'état est principalement de pouvoir spécifier des environnements et des évènements présentant des échéances. Typiquement, on peut intégrer dans le processus de prise de décision les dépendances temporelles qu'impliquent une affirmation du type : "la porte s'ouvrira dans cinq minutes et se fermera dans dix", à savoir qu'on peut définir une politique dépendant du temps car on a intégré ce dernier dans l'espace d'état. L'aspect stationnaire qui demeurait associé au SMDP est ici effacé car rendre le temps observable permet de décrire les phénomènes instationnaires (stationnaires à l'infini dans notre cas).

Le modèle avec dates comporte :
- Un état augmenté $\sigma \in \Sigma$ qui se décompose en :
  - Un espace d'état discret S
  - Un axe du temps $t \in \mathbb{R}$
- Un espace d'action discret A
- Une fonction de transition $P(\sigma'|\sigma,a)$ décomposable en $P(\sigma'|\sigma,a) = P(s'|\sigma,a) \cdot F(t'|\sigma,a)$
- Une fonction de récompense $r(\sigma', a, \sigma)$

La fonction $F(t'|\sigma,a)$ est une extension de la fonction $F(t|s,a)$ vue précédemment. La précédente décrivait la probabilité que l'action se termine en moins de t unités de temps ; la nouvelle décrit la probabilité que l'action se termine avant la date t. Il est important de bien noter la distinction entre durée et date. Par la suite, on s'efforcera de noter les dates t ou t' et les durées τ[2].

2.4 Le problème formalisé

Le problème auquel on s'intéresse est de résoudre l'équation de Bellman pour générer une politique du type $\pi(s,t)$. Cette équation s'écrit :

$$V^\pi(\sigma) = \sum_{s' \in S} \int_0^\infty (r(s', t+\tau, \pi(\sigma), \sigma) + \gamma^\tau V^\pi(\sigma')) \cdot F(\tau|\sigma, \pi(\sigma)) P(s'|\sigma, \pi(\sigma)) d\tau = L_\pi^t(V^\pi)(\sigma) \quad (7)$$

Dans ce cadre, l'algorithme d'itération de la valeur s'exprime sur la base des $(V_n)_{n \in \mathbb{N}}$:

$$V_0(\sigma) = 0$$

$$V_{n+1}(\sigma) = \max_{a \in A} \left\{ \sum_{s' \in S} \int_0^\infty (r(s', t+\tau, a, \sigma) + \gamma^\tau \cdot V_n(\sigma')) \cdot F(\tau|\sigma,a) P(s'|\sigma,a) d\tau \right\} = \max_{a \in A} L_a^t(V_n)(s,t)$$

[2] Il arrivera par ailleurs que, par commodité d'écriture, on confonde les notations $t + \tau$ et t'. On notera également indifféremment l'état augmenté σ ou (s,t).

En un état s, la politique optimale est une fonction du temps qui associe une action à chaque date t. La fonction de valeur $V_{n+1}(s,t)$ optimisée sur un coup en s est l'enveloppe du jeu de fonctions $(L_a^t(V_n)(s,t))_{a\in A}$. Si on note $V_a(s,t)$ la fonction de valeur associée à l'exécution de l'action a suivie de l'application de la politique π, alors optimiser π sur un coup revient à trouver l'enveloppe des $V_a(s,t)$ et les points où l'action correspondant à la fonction de valeur maximale change (les discontinuités possibles de la fonction r impliquent une fonction de valeur éventuellement discontinue). Le problème peut donc être vu comme un problème de découpage optimal, par état, de la droite temporelle. Chaque intervalle ainsi obtenu représente une période de décision durant laquelle l'action à entreprendre reste la même.

Dans le cadre qu'on vient d'introduire, nous avons développé trois approches différentes pour générer les politiques $\pi(s,t)$.

3 Le modèle exp-poly et une méthode de résolution formelle

L'idée qui a motivé le développement du modèle présenté ci-dessous est de parvenir à une résolution formelle de l'équation de Bellman. Le premier écueil auquel on peut s'attendre concerne le calcul de l'intégrale dans l'équation 7 ; ici, nous proposons d'introduire un cadre de modélisation riche dans lequel le calcul formel de cette intégrale est possible et simple.

3.1 Le modèle et ses hypothèses

Afin d'introduire ce modèle, nous avons défini la famille de fonctions exp-poly regroupant les fonctions qui, à t, associent le produit d'une exponentielle en t avec un polynôme en t[3]. Le modèle exp-poly repose sur les hypothèses suivantes que nous travaillons actuellement à élargir :

H1 : La fonction de récompense s'écrit : $r(\sigma', a, \sigma) = \sum_{j\in S} 1_j(s') \cdot rr_j(s,a) \cdot rt_j(t')$ avec s l'état de départ, j l'état d'arrivée et $rt_j(t')$ exp-poly, éventuellement discontinue.

H2 : Les durées et les probabilités de transition sont stationnaires, c'est-à-dire qu'on a : $\forall t \in \mathbb{R}^+,\ F(t'|\sigma,a) = F(\tau|s,a)$ et $P(s'|\sigma,a) = P(s'|s,a)$

H3 : La fonction $F(\tau|s,a)$ s'écrit comme :
- cas 1 : Une fonction constante par morceaux
- cas 2 : Une distribution à densité exp-poly
- cas 3 : Une combinaison de fonctions de répartition à densité Gaussienne.

[3]Eventuellement défini par morceaux. On notera au passage que les fonctions définies constantes par morceaux font partie de la famille des fonctions exp-poly.

L'hypothèse H1 nous permet de modéliser des récompenses différentes pour chaque transition vers un même état et de moduler ces récompenses en fonction du temps. Dans le cas du problème du taxi, cette modélisation est adaptée puisqu'elle nous permet de spécifier les éventuels différents points de dépot du passager indépendemment les uns des autres. L'hypothèse H2 est plus contraignante dans la mesure où elle impose que les durées ne varient pas avec la date : une transition aura toujours la même durée, quelle que soit la date à laquelle on entreprend l'action correspondante. Enfin, l'hypothèse H3 est une hypothèse de modélisation assez peu contraignante car avec les cas 1 et 2 on parvient à approcher n'importe quelle fonction positive croissante à valeurs dans $[0;1]$.

3.2 Recherche d'une expression formelle de la fonction de valeur

On peut alors montrer que la fonction de valeur $V^\pi(s,t)$ de toute politique π est de type exp-poly en t. Plus précisément, elle est définie par morceaux et les coefficients des polynômes des fonctions exp-poly sont calculables facilement grâce aux propriétés des fonctions exp-poly. Ainsi, pour stocker la fonction de valeur, il suffit de stocker les coefficients du polynôme et l'exposant de l'exponentielle sur chaque intervalle de définition de la fonction. La preuve a été effectuée pour les cas 1 et 2 pour l'expression de F ([Rac05]), le cas 3 est une piste non encore totalement explorée qui repose sur l'exploitation des propriétés des moments d'une fonction de répartition Gaussienne. En explicitant la preuve, on parvient notamment aux résultats suivants :

Dans le cas 1, pour un calcul itératif de la fonction de valeur, on maintient un cache de coefficientsdont la taille est de l'ordre de grandeur de $\big(deg(rt(t)) + 2\big) \cdot |S|^2 \cdot |A| \cdot n_{rF}$, avec n_{rF} le nombre moyen de morceaux de $rt_j(t')F(t'|\sigma,a)$ et $deg(rt(t))$ le degré maximum des polynômes des fonctions rt_j.
Dans le cas 2, on maintient un cache de coefficients dont la taille est de l'ordre de grandeur de $\big(deg(rt_j(t)) + 2\big) \cdot |S|^2 \cdot |A| \cdot n_r$, avec n_r le nombre moyen de morceaux de $rt_j(t')$. Ce calcul exact d'une fonction de valeur en (s,t) permet de construire une politique optimale en minimisant les calculs. La clé du processus est l'usage des propriétés des fonctions exp-poly, notamment le fait que la primitive d'une fonction exp-poly est une fonction exp-poly de même degré.

Lors des itérations de l'algorithme d'itération de la valeur, une étape clé de la détermination de V_{n+1} est la maximisation sur l'ensemble des $a \in A$. En un état discret s, on doit considérer toutes les fonctions exp-

poly par morceaux définies par les $(L_a^t(V_n)(s,t))_{a\in A}$ et en trouver l'enveloppe supérieure. Pour construire cette enveloppe, on procède de façon itérative : on définit la fonction enveloppe $V_{n+1}(s,t)$ comme égale à la première des $L_a^t(V_n)(s,t)$, puis on cherche les intersections avec la seconde des $L_a^t(V_n)(s,t)$ - pour trouver les intersections, si une résolution formelle n'est pas possible, on emploie une méthode de recherche numérique. On cherche alors la plus grande des deux fonctions sur chacun des intervalles ainsi définis (il suffit pour cela de tester un point quelconque car on sait qu'il n'y a pas d'inversion entre les fonctions sur ces intervalles). On remplace ainsi $V_{n+1}(s,t)$ par une nouvelle enveloppe, de valeur supérieure. Une fois qu'on a parcouru tout l'ensemble A, on dispose de la fonction de valeur $V_{n+1}(s,t)$. A chaque itération de l'algorithme d'itération de la valeur, on peut déterminer la politique courante définie sur S et sur chacun des morceaux de $V_{n+1}(s,t)$. On obtient ainsi une politique $\pi(s,t)$ donnant, pour chaque état s les plages de dates sur lesquelles chaque action à entreprendre est optimale.

La résolution formelle, si elle permet un traitement direct de l'équation de Bellman, présente toutefois deux faiblesses majeures. D'une part, là où on évaluait $|S| \cdot |A|$ valeurs lors d'une itération de la valeur classique, on stocke maintenant jusqu'à $|S| \cdot \big(deg(rt(t)) + 2\big) \cdot n$ ($n = n_{rF}$ ou n_r) fois plus de coefficients ce qui peut s'avérer lourd d'un point de vue informatique. Par ailleurs, lors de l'exécution de l'algorithme de maximisation de l'enveloppe des $(L_a^t(V_n)(s,t))_{a\in A}$, on calcule des intersections entre fonctions exp-poly, ce qui revient dans ce cas précis à trouver les racines d'autant de polynômes de degré $deg(rt(t))$. Cette opération peut-être particulièrement complexe et/ou coûteuse en temps de calcul car il s'agit d'un problème de maximisation continue sur une famille de fonctions discontinues et non-linéaires.
Pour ces raisons, nous conservons le cadre du modèle exp-poly pour de futurs usages mais il apparait nécessaire de simplifier l'algorithme de résolution pour pouvoir le rendre exploitable pour un calcul réel.

4 Une approche par discrétisation de la droite temporelle à pas fixe

Si le modèle précédent semble exploitable, ses hypothèses ne permettent pas de l'envisager dans un cadre plus général. Le calcul itératif d'une fonction de valeur formelle dans le cadre général est très fortement handicapé par le calcul des enveloppes et des intersections successives nécessaires à l'établissement de la suite des fonctions de valeur. Afin d'éviter les écueils évoqués précédemment, nous avons cherché à nous ramener à un cas discret, de façon similaire au modèle SMDP. Cependant, nous ne souhaitions ne pas trop perdre en optimalité. L'approche suivante a été motivée par la mise en place d'un algorithme de coopération et de coordination entre deux agents dans un cadre décentralisé et distribué, cependant l'idée de base s'applique indépendemment du nombre d'agents.

L'idée est de transformer la fonction F en une distribution sur une variable aléatoire de durée à valeurs discrètes. C'est lors de cette transformation que l'on perd en optimalité car on exclut les durées intermédiaires entre deux valeurs discrètes. On prend alors le plus grand commun diviseur (pgcd) des durées ainsi définies et on construit un espace de dates de décision discrétisé avec un tel pas jusqu'à la date du pseudo-horizon T_f.

On commence par transformer la distribution continue $F(\tau|s,a)$ en une distribution sur une variable de durée discrète. On obtient ainsi un jeu de toutes les durées de transition envisageables dans le cadre de notre problème. On prend alors le pgcd T de ces durées et on enrichit l'espace d'état avec une variable discrète t qui peut prendre les valeurs suivantes :
Avec $n = E(\frac{T_f}{T})$, $t \in \{\ 0, 1, \ldots, nT, (n+1)T\}$.

On redéfinit ainsi le modèle de récompense et de transition avec cet espace d'état discret enrichi, on ajoute la définition des récompenses qu'on a introduite avec le modèle SMDP et le problème final devient un problème MDP classique à résoudre. On peut remarquer notamment que les matrices de transition seront particulièrement creuses car, à partir d'un instant t_1, il n'y a qu'un petit nombre d'instants postérieurs atteignables en une action. On peut donc envisager diverses méthodes plus compactes de représentation et de résolution du problème (factorisation, décomposition).

Nous avons développé ce modèle plus en détail dans [Rac05]. Il s'avère qu'il est particulièrement utile pour définir des problèmes simples où une distribution discrète sur les durées suffit à décrire l'univers et à planifier. Cependant, la perte d'optimalité due à la discrétisation du domaine temporel n'est pas inéluctable comme on peut le voir à la section suivante.

5 La recherche de dates de décision par optimisation de l'erreur de Bellman

5.1 Présentation

Nos modèles précédents cherchent à construire des politiques étant donné une variable d'état t dont

les valeurs sont connues à l'avance ce qui rend son cardinal très grand. Cependant, l'expression de la solution finale ne nécessite pas la définition d'une action par unité de temps. Il est notamment beaucoup plus compact et utile de stocker une politique du type "en s, entre t_0 et t_1, faire a_1, entre t_1 et t_2 faire a_2, etc. jusqu'au pseudo horizon" que de spécifier l'action à faire à chaque instant. L'idée que nous développons dans cette partie est que si on trouve, pour chaque état discret, les dates de décision optimales - c'est-à-dire les dates exactes où l'action à entreprendre optimale change - alors on n'aura à rechercher les actions optimales qu'à ces dates là. L'approche que nous proposons ici réutilise des résultats des sections précédentes et les synthétise dans une méthode itérative de recherche des dates de décision optimales et de peuplement d'espaces de périodes de décision par état ; elle peut s'apparenter dans l'idée à l'approche de [Par98] pour le découpage des plages de valeurs de sortie dans le cadre de l'optimisation d'un MDP décomposé mais l'application et la mise en oeuvre en sont assez éloignées.

Cette approche semble particulièrement intéressante comparée aux deux précédentes car :
- Elle évite l'écueil du calcul intégral et de la maximisation sur une famille de $|A|$ fonctions de valeur (section 3.2),
- Elle évite l'explosion de l'espace d'état de la section 4 ou la multiplication des coefficients de la section 3.2,
- Elle renvoie des politiques compactes comme on l'a souhaité en section 1.

Il s'agit donc d'une réponse possible au problème tel qu'il a été défini à la section 2.4. Par la suite, on distinguera souvent l'ensemble de ces dates de décision de l'ensemble des états discrets du système, bien que tous fassent bien partie de l'espace d'état sur lequel est défini la politique. On commence par introduire quelques définitions :

Définition 1. *Une période de décision est un intervalle temporel sur lequel, pour un état donné, l'action à entreprendre est constante.*

Définition 2. *Une période de décision optimale est un intervalle temporel sur lequel, pour un état donné, l'action à entreprendre est toujours l'action optimale.*

Définition 3. *On note $\tilde{T}$ l'ensemble des dates de début de période de décision. Les éléments de $\tilde{T}$ sont les dates $\tilde{t}_i$.*

Définition 4. *On note $\tilde{T}_s$ l'ensemble des dates de début de période de décision pour l'état s. On définit un opérateur de masque M_s tel que : $\tilde{T}_s = \tilde{T}M_s$.*

Définition 5. *On note, pour tout $s \in S$, $(S, \tilde{T}M_s)$ la donnée de l'espace d'état discret et des $\tilde{t}_k$ utiles en s.*

Tout cela nous permet de redéfinir une politique comme :

Définition 6. *Une politique π est une fonction de $S \times \tilde{T}M_s$ dans A.*

On initialise l'algorithme en créant l'ensemble $\tilde{T}$. Initialement, cet ensemble ne contient que la date 0 et le pseudo-horizon. On se dote également d'opérateurs de masque M_s qui sont initialisés pour $\tilde{T}_s = \tilde{T}$. Relativement à un ensemble ($\tilde{T}$ ou $\tilde{T}_s$), on assimilera la date $\tilde{t}_k \in \tilde{T}_s$ de début de période de décision et la période elle-même ($[\tilde{t}_k; \tilde{t}_{k+1}]$ avec $\tilde{t}_k$ et $\tilde{t}_{k+1} \in \tilde{T}_s$). L'algorithme cherche à peupler les $\tilde{T}_s$ avec les dates de décision les plus utiles et uniquement celles-ci. On rajoute ainsi, à chaque itération, dans chaque $\tilde{T}_s$, la date de décision pour laquelle l'erreur de Bellman pour la décision actuelle est la plus grande. On définit ainsi une famille d'instants de décision limitée pour chaque état. L'objectif de notre algorithme est de trouver tous les $\tilde{T}_s$ tels que π soit ϵ-optimale.

5.2 L'algorithme de peuplement des ensembles de périodes de décision

Conformément au 2.3, on se dote de :
- Un espace d'état discret S,
- Un espace temporel initial $\tilde{T}$ peuplé avec toutes les échéances connues du problème,
- Un espace d'action A,
- Une fonction de transition $P(s'|s,a,t)$
- Une fonction de durée $F(t'|s,a,t)$
- Une fonction de récompense $r(a,s,t)$

L'algorithme procède alors en quatre étapes :

Première étape : génération de $\tilde{P}$ et $\tilde{r}$. La fonction $\tilde{P}(s', \tilde{t}'_j | a, s, \tilde{t}_k)$ décrit la probabilité que l'action a, entreprise en s pendant la période de décision $\tilde{t}_k$ de $\tilde{T}_s$, mène l'agent en s', pendant la période de décision $\tilde{t}'_j$ de $\tilde{T}_{s'}$. De même, $\tilde{r}$ représente l'espérance des récompenses associées à la transition $(\tilde{t}'_j, s', a, s, \tilde{t}_k)$. On a :

$$\tilde{P}(s', \tilde{t}'_j | a, s, \tilde{t}_k) = Pr(s'|a,s,t \in \tilde{t}_k) \cdot Pr(t' \in \tilde{t}'_j | a, s, t \in \tilde{t}_k) = P_1 \cdot P_2 \quad (8)$$

P_1 représente l'espérance, sur l'intervalle $[\tilde{t}_k; \tilde{t}_{k+1}]$, des probabilités d'arriver en s' à la fin de la transition. On suppose qu'on a une distribution uniforme sur la date de début entre $\tilde{t}_k$ et $\tilde{t}_{k+1}$, on peut donc écrire :

$$P_1 = Pr(s'|a,s,t \in \tilde{t}_k) = \frac{1}{\tilde{t}_{k+1} - \tilde{t}_k} \int_{\tilde{t}_k}^{\tilde{t}_{k+1}} P(s'|s,a,t)dt$$

De même, P_2 représente l'espérance sur $[\tilde{t}_k; \tilde{t}_{k+1}]$ des probabilités que l'action se finisse avant $\tilde{t}_{j+1} \in \tilde{T}_{s'}$ et

après $\tilde{t}_j \in \tilde{T}_{s'}$. On a donc :

$$P_2 = Pr(t' \in \tilde{t}'_j | a, s, t \in \tilde{t}_k)$$
$$= \frac{1}{\tilde{t}_{k+1} - \tilde{t}_k} \int_{\tilde{t}_k}^{\tilde{t}_{k+1}} \left(F(\tilde{t}_{j+1}|s,a,t) - F(\tilde{t}_j|s,a,t) \right) dt$$

De même :

$$\tilde{r}(a, s, \tilde{t}_k) = \frac{1}{\tilde{t}_{k+1} - \tilde{t}_k} \int_{\tilde{t}_k}^{\tilde{t}_{k+1}} r(a,s,t)dt \qquad (9)$$

On obtient ainsi un MDP $\tilde{M}$ composé de :
- Un espace d'état discret $S \times \{\tilde{T}_s / s \in S\}$
- Un espace d'action A
- Une fonction de transition $\tilde{P}$
- Une fonction de récompense $\tilde{r}$

Seconde étape : optimisation du MDP $\tilde{M}$. On résout l'équation de Bellman pour $\tilde{M}$ par une méthode classique (itération de la valeur, de la politique, ...). On obtient ainsi une politique $\pi(s, \tilde{t} \in \tilde{T}_s)$ et une fonction de valeur associée $V^\pi(s, \tilde{t})$. Par rapport à une résolution sans prise en compte du temps, on a rajouté $Card(\tilde{T}_s)$ états là où il n'y en avait qu'un. On a donc multiplié la taille de l'espace d'état par $E_s(Card(\tilde{T}_s))$. Sur un exemple simple à une récompense et deux échéances (une d'apparition de la récompense, une de disparition), on s'aperçoit que $Card(\tilde{T}_s)$ vaut au maximum 3 et que les trois valeurs sont utiles à la spécification de la politique, on utilise bien un espace d'état de taille minimale pour notre problème.

Troisième étape : Recherche des dates de décision optimales. L'idée de la troisième étape est de chercher, étant donné la politique courante, les dates en lesquelles on pourrait le mieux améliorer cette politique. Pour cela, on utilise l'erreur de Bellman en s : $BE_s(t)$. On considère, pour tout $s \in S$, la fonction $V^\pi(s,t)$, évaluation de la valeur de la politique π prolongée sur $S \times \mathbb{R}$. On a alors :

$$BE_s(t) = \max_{a \in A} \Big\{ r(s,a,t) + \sum_{s'} \int_t^\infty \gamma^{(t'-t)} V^\pi(s',t') F(t'|s,t,a) P(s'|s,t,a) dt' \Big\} - V^\pi(s,t) \qquad (10)$$

Cette erreur $BE_s(t)$ représente la quantité dont on peut améliorer la valeur de la politique π en s en fonction du temps en optimisant sur un coup. C'est une mesure de l'écart à l'optimum. Pour chaque état s, on cherche les triplets (s, t_s, ϵ_s) où ϵ_s est le maximum de $BE_s(t)$ atteint en t_s. On a ainsi, par état, l'instant où la décision courante est la plus "optimisable". La recherche des triplets (s, t_s, ϵ_s) est au coeur de notre algorithme et sera discutée plus loin (section 5.3).

Quatrième étape : Peuplement de $\tilde{T}$. Si ϵ_s est supérieur à un certain ϵ que l'on s'est fixé initialement, alors on ajoute t_s à $\tilde{T}_s$. On procède ainsi pour tous les s.

On recommence alors à l'étape 1. Lorsqu'on itère, à partir de la première itération, il faut insérer une phase de simplification des $\tilde{t}_s$ entre l'étape 2 et l'étape 3. En effet, si deux dates de décision consécutives $\tilde{t}_k$ et $\tilde{t}_{k+1}$ de $\tilde{T}_s$ ont la même action optimale, alors on peut supprimer de $\tilde{T}_s$ la plus tardive. On maintient ainsi, par état s, un cache de dates de décision minimal. On itère ainsi jusqu'à ce que tous les $\epsilon_s < \epsilon$ ou que l'on arrête les itérations (fonctionnement "anytime").

5.3 L'utilisation du modèle exp-poly pour la recherche de la meilleure date de décision

Le problème clé de notre algorithme est la résolution de l'étape 3, c'est-à-dire la recherche, par état, des dates de décision optimales par maximisation de l'erreur de Bellman. On cherche donc à maximiser sur t le terme de droite de l'équation 10. On a :

$$BE_s(t) = \max_{a \in A} \left\{ L_a^t(V^\pi)(s,t) \right\} - V^\pi(s,t) \qquad (11)$$

On cherche :

$$\max_{t \in \mathbb{R}} BE_s(t) = \max_{t \in \mathbb{R}} \max_{a \in A} \left\{ L_a^t(V^\pi)(s,t) - V^\pi(s,t) \right\}$$
$$= \max_{a \in A} \max_{t \in \mathbb{R}} \left\{ L_a^t(V^\pi)(s,t) - V^\pi(s,t) \right\}$$

Pour un s donné, on va donc chercher le maximum de $L_a^t(V^\pi)(s,t) - V^\pi(s,t)$ pour tous les a on prendra alors le maximum sur cette famille de maxima. Le problème se ramène donc à trouver le maximum sur t, à s et a fixés, de $L_a^t(V^\pi)(s,t) - V^\pi(s,t)$. En supposant cette expression dérivable par morceaux et en se plaçant sur chaque morceau, on cherche donc à résoudre l'équation :

$$\frac{\partial \left(L_a^t(V^\pi)(s,t) - V^\pi(s,t) \right)}{\partial t} = 0 \qquad (12)$$

On rappelle ici que :

$$L_a^t(V^\pi)(s,t) = r(s,a,t) + \sum_{s'} \int_t^\infty \gamma^{(t'-t)} V^\pi(s',t') F(t'|s,t,a) P(s'|s,t,a) dt'$$

Pour résoudre l'équation 12, on dispose de deux possibilités. La première consiste à tenter une résolution numérique. Nos travaux actuels cherchent entre autres à borner l'erreur résultante des méthodes numériques.

Une autre possibilité est de reprendre le modèle exp-poly. En effet, dans ce modèle, V^π est solution de l'équation 7.

L'expression de $L^t_\pi(V^\pi)(s,t)$ peut se calculer rapidement de façon récursive grâce au calcul formel des coefficients des polynômes des fonctions exp-poly : la primitive d'une fonction exp-poly est une fonction exp-poly de même degré et on peut exprimer les coefficients de la primitive en fonction de la fonction d'origine. On dispose ainsi d'une expression de $L^t_a(V^\pi)(s,t) - V^\pi(s,t)$ dérivable par morceaux, et il reste à résoudre l'équation 12 (si la résolution formelle n'est pas possible, une méthode numérique peut être appliquée à la recherche de ce maximum).
Contrairement à l'approche formelle de la section 3.2 où chaque intersection trouvée sert à redéfinir les morceaux de la fonction de valeur pour être réinjecté dans l'équation de Bellman à l'itération suivante, la valeur que l'on trouve ici est la valeur finale que l'on va utiliser pour peupler $\tilde{T}_s$, on n'a donc pas le souci de la perte d'optimalité due à la propagation des erreurs numériques.

On trouve ainsi $|A|$ maxima des $L^t_a(V^\pi)(s,t) - V^\pi(s,t)$. L'erreur de Bellman en s correspond alors au plus grand d'entre eux. On dispose ainsi du triplet (s, t_s, ϵ_s).

5.4 L'adaptation de l'algorithme à un modèle discret

Enfin, par souci de simplicité et d'applicabilité, on peut restreindre le champ d'application de notre algorithme afin de faciliter les résolutions : on peut s'intéresser directement à des fonction P, F et r définies sur une variable t discrète. On peut éventuellement également considérer une fonction F déterministe (la part d'incertain dans les transitions est conservée dans P). On pourra ainsi obtenir des politiques simples et applicables dépendant du temps. Les premiers cas de test considérés sont d'ailleurs envisagés dans ce cadre.

6 Conclusion

Au final, pour pouvoir planifier dans l'incertain en présence d'un univers instationnaire (stationnaire à l'infini), nous avons introduit un modèle rendant la variable temporelle observable par l'agent. Sur la base de ce modèle et de la réécriture de l'équation de Bellman, nous avons proposé trois approches différentes pour la génération de politiques du type $\pi(s,t)$. La résolution formelle a montré ses limites du point de vue du calcul de l'optimisation de la fonction de valeur. L'approche par discrétisation à pas discret semble une piste intéressante pour les problèmes simples mais elle ne résout que partiellement le problème de l'explosion de l'espace d'état. Enfin, l'approche par recherche des dates de décision optimales, tout en reprenant les résultats développés dans les deux approches précédentes, semble répondre au problème initial de façon adéquate : il ne semble pas y avoir d'écueil calculatoire et l'espace d'état envisagé en permanence est minimal.

L'implémentation de l'algorithme est en cours et les premiers résultats de résolution "à la main" sont encourageants. Nous travaillons également sur l'extension de l'étape 3 à une résolution ne dépendant pas du modèle de P, F et r. A terme, l'objectif est d'inclure cet algorithme dans une boucle de niveau supérieur déjà développée permettant à deux agents de communiquer leur stratégie de façon compacte afin de parvenir à un plan global en planifiant chacun de son côté (donc de façon complètement décentralisée et distribuée).

Références

[Bel57] R.E. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, New Jersey, 1957.

[BM05] A. Beynier and A.-I. Mouaddib. Processus décisionnels de markov décentralisés et interactifs avec des contraintes temporelles. *Journal électronique d'intelligence artificielle*, 4(22), 2005.

[Bou96] C. Boutilier. Planning, learning and coordination in multiagent decision processes. In *Theoretical Aspects of Rationality and Knowledge*, pages 195–201, 1996.

[CC03] I. Chades and F. Charpillet. Modèles de conception des sma coopératifs par planification réactive. In *2e Journées Francophones Modèles Formels de l'Interaction*, 2003.

[Die98] T.G. Dietterich. The MAXQ method for hierarchical reinforcement learning. In *Proc. 15th International Conf. on Machine Learning*, pages 118–126. Morgan Kaufmann, San Francisco, CA, 1998.

[Par98] R. Parr. Flexible decomposition algorithms for weakly coupled markov decision problems, 1998.

[Put94] M.L. Puterman. *Markov Decision Processes*. John Wiley & Sons, Inc, 1994.

[Rac05] E. Rachelson. Coordination multi-robots terrestre et aérien - rapport de m2r. Technical report, ONERA-DCSD Toulouse, 2005.

[TC04] A.I. Tavares and M.F.M. Campos. Balancing coordination and synchronization cost in cooperative situated multi-agent systems with imperfect communication. In *ECAI*, 2004.