XMDP : un modèle de planification temporelle dans l'incertain à actions paramétriques

Emmanuel Rachelson[1], Florent Teichteil[1], Frédérick Garcia[2]

[1] ONERA-DCSD
2, Avenue E. Belin ; 31055 Toulouse
emmanuel.rachelson, florent.teichteil@onera.fr

[2] INRA-BIA
Chemin de Borde Rouge ; 31326 Castanet-Tolosan
fgarcia@tlse.toulouse.inra.fr

Résumé : Certains problèmes de décision impliquent de choisir à la fois des actions à entreprendre mais également des paramètres à affecter à ces actions. Par exemple, l'action "avancer" nécessite souvent d'y associer une distance. Dans le cadre de la décision dans l'incertain, on propose d'étendre le modèle MDP pour prendre en compte des actions paramétriques dont le paramètre est une variable de décision. On s'attache à établir les équations d'optimalité pour ces MDP paramétriques et on prolonge les résultats connus pour les MDP classiques. La variable temporelle a une place spéciale dans ce modèle, on détaillera ses propriétés et on les mettra en lumière des travaux précédents en planification temporelle dans l'incertain et en MDP à espaces d'état hybrides.

1 Introduction

Imaginons un robot terrestre devant planifier sa progression au milieu d'un feu de forêt où les routes sont praticables ou non, selon l'heure de la journée et la progression de l'incendie, et où les résultats et les durées de ses actions sont incertains. Le plan construit doit prendre en compte l'incertitude sur les résultats des actions modélisées mais il arrive que le jeu d'actions discrètes comme "attendre cinq minutes", "avancer d'un mètre", etc. ne soit pas adapté au problème : notre agent peut trouver un plan optimal vis-à-vis de ce jeu d'actions mais ce plan sera inadapté par rapport au problème réel. En revanche, supposons que l'on laisse l'agent décider des paramètres de distance, de durée, etc. qu'il affecte à ses actions. Le problème de décision devient plus complexe mais en retour la stratégie générée sera plus adaptée au problème en question. Dans cet article, nous présentons une manière d'introduire des actions paramétriques dans le cadre des Processus Décisionnels de Markov (MDP). On s'attache à fournir une définition rigoureuse du cadre, on met en lumière l'importance et les caractéristiques de la variable temporelle dans le modèle et on fournit les preuves d'équivalence entre critère et équation d'optimalité de Bellman.

Le cadre des MDP est un cadre classique de représentation des problèmes de décision dans l'incertain. Pour le type de problèmes évoqué ci-dessus, de nombreux travaux ont été effectués dans le cadre de la planification avec ressources continues (Bresina *et al.*, 2002), de la planification temporelle (Boyan & Littman, 2001; Younes & Simmons, 2004) et pour la prise en compte de variables discrètes et continues dans l'état (Guestrin *et al.*, 2004; Hauskrecht & Kveton, 2006). Cependant, peu de travaux portent sur les actions continues bien que le problème ait été mentionné comme important dans la conclusion de (Feng *et al.*, 2004). Considérer un espace d'action continu de dimension infinie n'a que peu de sens physique en général ; en revanche, un espace d'actions paramétriques est un cas courant dans les problèmes réels. Les travaux présentés étendent le cadre MDP à ce type d'action avec un accent particulier sur le paramètre de durée.

Ces travaux s'inscrivent dans le lien fort qui existe entre théorie de la décision et commande optimale (Bertsekas & Shreve, 1996). Comme en commande optimale, nous considérons un espace de commande continu, toutefois la similitude s'arrête là : contrairement aux problèmes d'automatique, notre cadre traite de problèmes de décision séquentielle où les étapes de décision se succèdent avec des actions prises dans un espace d'actions continu, tandis qu'en commande optimale, c'est l'asservissement continu d'un ou plusieurs paramètres qui est étudié. De plus, le modèle des actions paramétriques prend en compte des étapes de décision à dates aléatoires, ce qui le distingue à la fois des modèles classiques de décision dans l'incertain et des modèles d'automatique. Cet aspect clé génère la complexité des preuves présentées ici mais permet surtout de planifier en environnement incertain instationnaire avec une variable temporelle continue observable comme on le verra en section 2.

Nous commençons par définir le modèle des MDP à actions paramétriques. Nous précisons notamment les hypothèses sur les espaces d'états et d'actions à composantes hybrides (continues et discrètes) , sur la durée minimale d'action α et sur la propriété de semi-continuité supérieure du modèle de récompense vis-à-vis du ou des paramètres. Dans ce cadre, nous étendrons les notions de politique, de critère et d'optimalité. Puis, à la section 3 nous prouverons l'équivalence entre critère et équation de Bellman. Enfin nous présenterons des résultats plus généraux en section 4 et comparerons le modèle proposé aux travaux existants en section 5 pour conclure en section 6.

2 MDP et actions paramétriques

2.1 Présentation du modèle

Nous utiliserons par la suite des notations MDP classiques (Puterman, 1994) et décrirons un MDP comme un quintuplet $< S, A, P, r, T >$ avec un espace d'états S, un espace d'actions A, un modèle de transitions P associant à chaque transition (s', a, s) une probabilité, un modèle de récompense r et un ensemble d'étapes de décision T. Dans le cadre général des MDP à horizon infini, T est isomorphe à $\mathbb{N}$.

Définition 1 (MDP à actions paramétriques)
Un MDP à actions paramétriques est un sextuplet $< S, A(X), p, r, T >$ *où :*
S *est un espace d'états Borélien décrivant des variables d'état continues ou discrètes.* A *est un espace d'actions décrivant un jeu fini d'actions* $a_i(x)$ *avec* x *un vecteur de paramètres prenant ses valeurs dans un espace vectoriel* X*. De ce fait, l'espace d'actions* $A(X)$ *du problème est un espace continu.* p *est une densité de probabilités* $p(s'|s, a(x))$*.* r *est une fonction de récompense* $r(s, a(x))$*.* T *est un ensemble de périodes de décision.*

Afin de simplifier les notations, nous ne considérerons qu'un espace à une dimension pour X et ne prendrons donc en compte que des paramètres de durée d'action $x = \tau$. Ainsi $X = \mathbb{R}$ (en fait $X = \mathbb{R}^{+*}$ pour des considérations physiques : on n'a pas de durée négative). Ce choix - fait pour alléger les démonstrations - est justifié par les deux points suivants : tout d'abord nous verrons que le paramètre τ a une importance particulière vis-à-vis de l'ensemble T et du critère. D'autre part, le traitement des autres paramètres est similaire et plus simple et nous voulons éviter d'obscurcir les démonstrations par d'autres paramètres d'action que les durées. Nous considérerons donc un espace d'état hybride $s \in S$ comportant une variable temporelle continue $t \in \mathbb{R}$ (le temps est bien une variable d'état, mais par commodité, on notera l'état (s, t)). Cette écriture met en évidence l'importance de la variable temporelle. Cette distinction entre s et t au sein d'une variable d'état plus générale (que l'on pourrait écrire $\sigma = (s, t)$ comme dans (Rachelson *et al.*, 2006)) est surtout faite pour mettre en évidence les particularités des preuves données en section 3. On peut noter que pour des variables discrètes, la fonction $p()$ est une distribution discrète et que les intégrales sur $p()$ sont équivalentes à des sommes sur la variable discrète.

Un exemple jouet qui illustre ce modèle avec un espace d'états discrets est présenté figure 1. Il s'agit d'un jeu où le joueur doit amener une balle d'une case de départ à une case d'arrivée sur le plateau de jeu. Le joueur doit se dépêcher car la balle fond et le fond de certaines cases s'escamote avec une probabilité connue dépendant de la date. Le joueur dispose de cinq actions : "attendre τ secondes" ou pousser la balle dans l'une des quatre directions. L'état est décrit par la position de la balle et la date courante.

On peut remarquer que le fait de considérer à la fois un paramètre de durée et une incertitude sur les résultats d'action implique de considérer un temps continu et - plus important - de considérer des périodes de décision dont la date est dans $\mathbb{R}$ et non plus dans $\mathbb{N}$. C'est ce qui rend le paramètre τ unique. Nous verrons à la section suivante comment cette particularité affecte les équations d'optimalité. Pour l'instant, notons simplement δ le numéro de la période de décision et t_δ la variable aléatoire à valeurs réelles représentant la date à laquelle on prend la décision δ.

2.2 Politiques et critère

On définit une *règle de décision* (Puterman, 1994) à la période de décision δ comme l'application : $d_\delta : \left\{ \begin{array}{ccc} S \times \mathbb{R} & \rightarrow & A \times X \\ s, t & \mapsto & a, x \end{array} \right.$. d_δ indique l'action paramétrique à entreprendre en (s, t) à la période de décision δ. Une *politique* est définie comme un

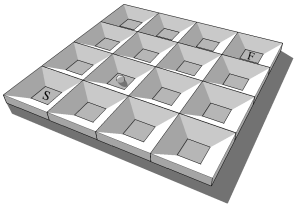
Objectif : Amener la balle de S à F.		Actions : attendre(τ), pousser dans une direction.
— La balle fond. — Le fond des cases est escamotable.		Les résultats et durées d'action sont incertains.

TABLE 1 – Exemple illustratif : la "balle fondante"

jeu de règles de décision (une par δ) et l'on considère, comme dans (Puterman, 1994), l'ensemble $\mathcal{D}$ des politiques Markoviennes déterministes stationnaires (par rapport à δ).

Afin de définir des politiques optimales pour notre problème, il faut préciser un critère. Le modèle SMDP (Puterman, 1994) propose une extension des MDP aux modèles stationnaires à temps continu. Il est décrit par des espaces S et A discrets, un modèle de transition $P(s'|s,a)$ et une fonction $F(t|s,a)$ indiquant la probabilité de durée des transitions. Le critère γ-pondéré pour les SMDP intègre alors les récompenses espérées pour toutes les durées de transition possibles. Similairement au critère γ-pondéré pour les SMDP, nous définissons le critère γ-pondéré pour les XMDP comme l'espérance de la somme des récompenses pondérées successives, obtenues en appliquant une politique π à partir d'un état initial (s,t) :

$$V_\gamma^\pi(s,t) = E_{(s,t)}^\pi \left\{ \sum_{\delta=0}^{\infty} \gamma^{t_\delta - t} r_\pi(s_\delta, t_\delta) \right\} \tag{1}$$

Afin de s'assurer de la convergence de cette série, on fait les hypothèses suivantes :
— $|r\left((s,t),a(\tau)\right)|$ est borné par M,
— $\exists \alpha / \forall \delta \in T, \quad t_{\delta+1} - t_\delta \geq \alpha > 0$, avec α la plus petite durée d'action,
— $\gamma < 1$.
Le facteur de pondération γ^t assure alors la convergence de la série. Physiquement, on peut le voir comme la probabilité que le processus continue après un temps t. Avec ces hypothèses, il est aisé de voir que pour tout $(s,t) \in S \times \mathbb{R}$:

$$|V_\gamma^\pi(s,t)| < \frac{M}{1-\gamma^\alpha} \tag{2}$$

Nous admettrons ici que l'ensemble $\mathcal{V}$ des fonctions de valeur (fonctions de $S \times \mathbb{R}$ dans $\mathbb{R}$) est un espace métrique complet pour la norme infini $\|V\|_\infty = \sup_{(s,t)\in S\times\mathbb{R}} V(s,t)$. Une politique optimale π^* est alors définie comme vérifiant $V_\gamma^{\pi^*} = \sup_{\pi\in\mathcal{D}} V_\gamma^\pi$. L'existence d'une telle politique dont la valeur atteint le $\sup$ est garantie par l'hypothèse de semi-continuité supérieure de r. A partir d'ici, on omettra l'indice γ sur V.

On cherche alors une manière de caractériser la stratégie optimale. Un MDP est classiquement résolu par programmation dynamique (Bellman, 1957) ou linéaire (Guestrin *et al.*, 2004) sur la base de l'équation d'optimalité de Bellman. Suivant ce principe, nous nous concentrons sur l'équation d'optimalité et prouvons l'équivalence entre le critère que nous avons introduit et une équation de Bellman usuelle.

3 Equation d'optimalité

Dans cette partie nous établissons et démontrons l'existence d'une équation d'optimalité pour les XMDP. Nous introduisons en premier lieu l'opérateur L^π d'évaluation d'une politique, puis redéfinissons l'opérateur de programmation dynamique L pour les XMDP et prouvons que V^* est l'unique solution de $V = LV$.

3.1 Evaluation de la politique

Définition 2 (opérateur L^π)
L'opérateur L^π associe, à tout élément V de $\mathcal{V}$ la fonction de valeur :

$$L^\pi V(s,t) = r(s,t,\pi(s,t)) + \int_{\substack{t' \in \mathbb{R} \\ s' \in S}} \gamma^{t'-t} p(s',t'|s,t,\pi(s,t)) V(s',t') ds' dt' \qquad (3)$$

On note que pour des actions non paramétriques et dans un espace d'états discret, $p()$ est une densité de probabilités discrète, les intégrales deviennent des sommes et le L^π ci-dessus devient égal à l'opérateur L^π pour les MDP classiques. Cet opérateur représente le gain si l'on applique π sur un coup une fois avant d'obtenir V. Nous allons prouver que cet opérateur peut-être utilisé pour caractériser la valeur de la politique π.

Proposition 1 (Evaluation de la politique)
Soit π une politique de $\mathcal{D}$. $V = V^\pi$ est l'unique solution de l'équation $L^\pi V = V$.

Preuve Dans tout ce qui suit, la notation $E^\pi_{a,b,c}$ décrit une espérance conditionnelle par rapport à π, sachant les valeurs des variables aléatoires a, b and c. Plus précisément, $E^\pi_{a,b,c}(f(a,b,c,d,e))$ représente l'espérance de f calculée par rapport à d et e et est, de fait, une fonction de a, b et c.

Notre état initial est $(s_0,t_0) = (s,t)$:

$$\begin{aligned} V^\pi(s,t) &= E^\pi_{s_0,t_0}\left\{\sum_{\delta=0}^{\infty} \gamma^{t_\delta - t} r_\pi(s_\delta,t_\delta)\right\} \\ &= r_\pi(s,t) + E^\pi_{s_0,t_0}\left\{\sum_{\delta=1}^{\infty} \gamma^{t_\delta - t} r_\pi(s_\delta,t_\delta)\right\} \\ &= r_\pi(s,t) + E^\pi_{s_0,t_0}\left\{E^\pi_{\substack{s_0,t_0 \\ s_1,t_1}}\left(\sum_{\delta=1}^{\infty} \gamma^{t_\delta - t} r_\pi(s_\delta,t_\delta)\right)\right\} \end{aligned}$$

Le terme entre accolades est calculé par rapport aux variables $(s_i,t_i)_{i=2\ldots\infty}$. On l'intègre alors sur les variables (s_1,t_1) afin d'obtenir l'espérance générale sachant (s_0,t_0). Nous développons le premier calcul d'espérance en notant $(s_1,t_1) = (s',t')$:

$$V^\pi(s,t) = r_\pi(s,t) + \int_{\substack{t' \in \mathbb{R} \\ s' \in S}} E^\pi_{\substack{s_0,t_0 \\ s_1,t_1}}\left(\sum_{\delta=1}^{\infty} \gamma^{t_\delta - t} r_\pi(s_\delta,t_\delta)\right) p_\pi(s',t'|s,t) ds' dt'$$

$$V^{\pi}(s,t) = r_{\pi}(s,t) + \int_{\substack{t' \in \mathbb{R} \\ s' \in S}} \gamma^{t'-t} p_{\pi}(s',t'|s,t) \cdot E^{\pi}_{\substack{s_0,t_0 \\ s_1,t_1}} \left(\sum_{\delta=1}^{\infty} \gamma^{t_\delta - t'} r_{\pi}(s_\delta, t_\delta) \right) ds' dt'$$

L'expression à l'intérieur du $E^{\pi}_{s_0,t_0,s_1,t_1}()$ dépend des variables aléatoires (s_i, t_i) pour $i \geq 2$. La propriété de Markov sur les probabilités $p()$ nous permet d'écrire que cette espérance ne dépend que des variables aléatoires (s_1, t_1) et donc que :

$$E^{\pi}_{\substack{s_0,t_0 \\ s_1,t_1}} \left(\sum_{\delta=1}^{\infty} \gamma^{t_\delta - t} r_{\pi}(s_\delta, t_\delta) \right) = V^{\pi}(s', t')$$

Et on a : $$V^{\pi}(s,t) = L^{\pi} V^{\pi}(s,t) \tag{4}$$

Cette solution est unique car L^{π} est contractante sur $\mathcal{V}$ et l'on peut utiliser le théorème du point fixe de Banach (la preuve que L^{π} est contractante sur $\mathcal{V}$ est similaire à celle que l'on propose pour l'opérateur L dans la section suivante). □

3.2 Operateur de Bellman

Définition 3 (opérateur L)
L'opérateur de Bellman de programmation dynamique L associe à tout élément V de $\mathcal{V}$ la fonction de valeur : $LV = \sup_{\pi \in \mathcal{D}} \{L^{\pi} V\}$

$$LV(s,t) = \sup_{\pi \in \mathcal{D}} \left\{ r_{\pi}(s,t) + \int_{\substack{t' \in \mathbb{R} \\ s' \in S}} \gamma^{t'-t} p_{\pi}(s',t'|s,t) V(s',t') ds' dt' \right\} \tag{5}$$

Cet opérateur représente l'optimisation sur un coup de la politique courante. Nous allons prouver que L permet de définir une équation d'optimalité équivalente à l'expression du critère γ-pondéré (equation 1). On peut remarquer que la semi-continuité supérieure des fonctions de récompense par rapport au modèle de récompense garantit qu'une telle borne supérieure existe et est atteinte dans l'équation 5.

Proposition 2 (équation de Bellman)
Pour un XMDP et un critère γ-pondéré, la fonction de valeur optimale est l'unique solution de l'équation $V = LV$.

Preuve La preuve ci-dessous adapte la preuve classique aux hypothèses plus générales présentées précédemment et résout les points dûrs soulevés. Notamment l'utilisation d'espaces d'états hybrides, d'espaces d'actions paramétriques et de fonctions de récompense semi-continues. Le raisonnement se fait en trois étapes :

1. Nous prouvons d'abord que si $V \geq LV$ alors $V \geq V^*$,
2. Puis on montre que si $V \leq LV$ alors $V \leq V^*$,
3. Enfin on montre que la solution de $V = LV$ est unique.

Supposons qu'il existe V tel que $V \geq LV$. Alors, avec π une politique de $\mathcal{D}$, on a : $V \geq \sup_{\pi \in \mathcal{D}} \{L^{\pi} V\} \geq L^{\pi} V$. L^{π} étant positive, on a récursivement : $V \geq L^{\pi} V \geq$

$L^\pi L^\pi V \ldots \geq L^{\pi(n+1)}V$. On souhaite trouver N tel que $\forall n \geq N, \quad L^{\pi(n+1)}V - V \geq 0$.
$L^{\pi(n+1)}V$ décrit l'application de π, $n+1$ fois avant de recevoir V.

$$L^{\pi(n+1)}V = r_\pi(s_0,t_0) + E^\pi_{s_0,t_0}\left(\gamma^{t_1-t_0} r_\pi(s_1,t_1) \; + E^\pi_{s_1,t_1}\left(\gamma^{t_2-t_0} r_\pi(s_2,t_2) + \ldots + E^\pi_{s_{n-1},t_{n-1}}\left(\gamma^{t_n-t_0} r_\pi(s_n,t_n) \; + E^\pi_{s_n,t_n}\left(\gamma^{t_{n+1}-t_0} V(s_{n+1},t_{n+1})\right)\right)\ldots\right)\right)$$

$$V^\pi = r_\pi(s_0,t_0) + E^\pi_{s_0,t_0}\left(\gamma^{t_1-t_0} r_\pi(s_1,t_1) \; + E^\pi_{s_1,t_1}\left(\gamma^{t_2-t_0} r_\pi(s_2,t_2) + \ldots + E^\pi_{s_{n-1},t_{n-1}}\left(\gamma^{t_n-t_0} r_\pi(s_n,t_n) \; + E^\pi_{s_n,t_n}\left(\sum_{\delta=n+1}^{\infty} \gamma^{t_\delta-t_0} r_\pi(s_\delta,t_\delta)\right)\right)\ldots\right)\right)$$

En écrivant $L^{\pi(n+1)}V - V^\pi$ on peut rassembler les deux expressions ci-dessus en une fonction d'espérance globale sur les $(s_i,t_i)_{i=0\ldots\infty}$. Tous les premiers termes s'annulent alors un à un et on peut écrire :

$$L^{\pi(n+1)}V - V^\pi = E^\pi_{\substack{(s_i,t_i)\\ i=0\ldots n}}\left(\gamma^{t_{n+1}-t_0} V(s_{n+1},t_{n+1}) \; - \sum_{\delta=n+1}^{\infty} \gamma^{t_\delta-t_0} r_\pi(s_\delta,t_\delta)\right)$$

et ainsi : $L^{\pi(n+1)}V - V^\pi = E^\pi_{\substack{(s_i,t_i)\\ i=0\ldots n}}\left(\gamma^{t_{n+1}-t_0} V(s_{n+1},t_{n+1})\right) - E^\pi_{\substack{(s_i,t_i)\\ i=0\ldots n}}\left(\sum_{\delta=n+1}^{\infty} \gamma^{t_\delta-t_0} r_\pi(s_\delta,t_\delta)\right)$

On écrit : $L^{\pi(n+1)}V - V^\pi = q_n - r_n$.

On a $\gamma < 1$, $r() < M$ et $\forall n \in \mathbb{N}, \quad t_{n+1} - t_n \geq \alpha > 0$; on sait donc que $\|V\|$ est bornée (équation 2) et on a : $E^\pi_{\substack{(s_i,t_i)\\ i=0\ldots n}}\left(\gamma^{t_{n+1}-t_0} V(s_{n+1},t_{n+1})\right) \leq \gamma^{(n+1)\alpha}\|V\|$.

On peut donc écrire $\lim_{n\to\infty} q_n = 0$.

Par ailleurs, r_n est le reste d'une série convergente. Donc on a : $\lim_{n\to\infty} r_n = 0$.

Et donc $\lim_{n\to\infty} L^{\pi(n+1)}V - V^\pi = 0$. On avait $V \geq L^{\pi(n+1)}V$, donc $V - V^\pi \geq L^{\pi(n+1)}V - V^\pi$. Le terme de gauche ne dépend pas de n et la limite du terme de droite est nulle, on peut donc écrire : $V - V^\pi \geq 0$.

Ceci étant vrai pour tous les $\pi \in \mathcal{D}$, c'est vrai pour π^* et : $V \geq LV \Rightarrow V \geq V^*$
Avec un raisonnement similaire, on peut montrer que si $\pi' = \arg\sup_{\pi\in\mathcal{D}} L^\pi V$ et $V \leq LV$, alors $V \leq L^{\pi'(n+1)}V$. Donc $V - V^{\pi'} \leq L^{\pi'(n+1)}V - V^{\pi'}$ et ainsi $V - V^{\pi'} \leq 0$.

Comme $V^{\pi'} \leq V^*$, on a : $V \leq LV \Rightarrow V \leq V^*$

Les deux assertions précédentes indiquent que s'il existe une solution à $V = LV$ alors cette solution est égale à V^*. Afin de terminer la preuve de la proposition, il faut montrer qu'il existe toujours une solution à $V = LV$.

$\mathcal{V}$ est un espace métrique, complet pour la norme infini $\|V\|_\infty = \sup_{(s,t)\in S\times\mathbb{R}} V(s,t)$ (Bertsekas & Shreve, 1996). Si l'on montre que L est contractante sur $\mathcal{V}$ alors on pourra appliquer le théorème du point fixe de Banach.

Soient U et V deux éléments de $\mathcal{V}$ tels que $LV \geq LU$. Soit (a^*, τ^*) la solution de :

$$a^*(\tau^*) \quad = \quad \underset{a(\tau)\in A(\mathbb{R})}{argsup} \left\{ r(s,t,a(\tau)) \quad + \int\limits_{\substack{t'\in\mathbb{R}\\ s'\in S}} \gamma^{t'-t} p_{a(\tau)}(s',t'|s,t) V(s',t') ds'dt' \right\}$$

(a^*, τ^*) existe grace à la semi-continuité du modèle de récompense qui garantit que, même à un point de discontinuité de la fonction de récompense, la valeur supérieure est atteignable. On sait que $\forall (s,t) \in S \times \mathbb{R}, \quad |LV(s,t) - LU(s,t)| = LV(s,t) - LU(s,t)$.

$$LV(s,t) - LU(s,t) \leq r(s,t,a^*(\tau^*)) + \int\limits_{\substack{t'\in\mathbb{R}\\ s'\in S}} \gamma^{t'-t} p_{a^*(\tau^*)}(s',t'|s,t) V(s',t') ds'dt' -$$
$$r(s,t,a^*(\tau^*)) - \int\limits_{\substack{t'\in\mathbb{R}\\ s'\in S}} \gamma^{t'-t} p_{a^*(\tau^*)}(s',t'|s,t) U(s',t') ds'dt'$$

$$LV(s,t) - LU(s,t) \leq \int\limits_{\substack{t'\in\mathbb{R}\\ s'\in S}} \gamma^{t'-t} p_{a^*(\tau^*)}(s',t'|s,t) \left(V(s',t') - U(s',t')\right) ds'dt'$$

On a : $\begin{cases} V(s,t) - U(s,t) \leq \|V - U\| \\ t' - t \geq \alpha > 0 \\ p(s',t'|s,t,a(\tau)) \leq 1 \\ \gamma < 1 \end{cases}$, on peut donc écrire :

$$LV(s,t) - LU(s,t) \leq \|V - U\| \cdot \gamma^\alpha, \text{ et donc :}$$
$$\|LV - LU\| \leq \|V - U\| \cdot \gamma^\alpha$$

Comme $\gamma^\alpha < 1$, l'équation précédente prouve que L est contractante sur $\mathcal{V}$. Le théorème du point fixe de Banach nous permet alors d'affirmer qu'il existe un unique point fixe $V' \in \mathcal{V}$ à l'opérateur L tel que $V' = LV'$.

Les résultats précédents nous permettent de conclure que sous les hypothèses générales mentionnées plus haut, l'équation $LV = V$ a une unique solution et que cette solution est égale à V^*, la fonction de valeur optimale vis-à-vis du critère γ-pondéré. □

Nous avons donc prouvé, sous certaines hypothèses, que l'équation d'optimalité de Bellman étendue au cadre XMDP reste valable.

$$LV = V \Rightarrow V = V^* \tag{6}$$

On peut réécrire l'équation de Bellman de façon à la rendre exploitable pour les algorithmes de programmation dynamique tels que l'itération de la valeur ou de la politique :

$$LV(s,t) = \max_{a \in A} \sup_{\tau \in \mathbb{R}^+} \left\{ r(s,t,a) + \int_{\substack{t' \in \mathbb{R} \\ s' \in S}} \gamma^{t'-t} p(s',t'|s,t,a(\tau)) V(s',t') ds' dt' \right\} \tag{7}$$

En utilisant cette formulation, on alterne une phase d'optimisation sur τ de la valeur de chaque action qui fournit la valeur optimale du paramètre de l'action, et une phase de choix parmi l'ensemble des action possibles (associées à leurs paramètres optimaux).

On remarque que si l'espace d'états est discret, toutes les densités de probabilité sont discrètes et les intégrales deviennent des sommes. Si l'espace des paramètres est également discret, alors en renumérotant les actions de l'espace d'actions, le $\sup$ devient un $\max$ et l'équation de Bellman ci-dessus (equation 6) devient l'équation standard caractérisant les solutions des MDP classiques. On conclut ainsi que le modèle XMDP et son équation d'optimalité incluent et généralisent les résultats pour les MDP classiques.

4 Généralisation

4.1 La spécificité de la variable temporelle

Les raisonnements précédents illustrent le fait qu'il fille différencier trois significations différentes de la variable temporelle dans le cadre XMDP :

— C'est l'indice δ de l'étape de décision, le temps de la chaîne de Markov induite par le XMDP et la politique associée,
— C'est la variable aléatoire t_δ caractérisant la date à laquelle chaque décision est prise. La principale différence avec les modèles classiques réside dans le fait qu'en général, on écrit implicitement $\delta = t_\delta$ et qu'alors t_δ est déterministe.
— C'est enfin le paramètre de durée τ, indiquant combien de temps on consacre à une action.

Des approches de traitement plus spécifiquement dédiées à la variable temporelle ont été proposées dans (Rachelson *et al.*, 2006), on peut notamment citer l'utilisation des techniques de discrétisation pour déterminer les intervalles temporels de décision optimaux pour un problème instationnaire à temps continu (algorithme SMDP+).

4.2 Espace des paramètres généralisé

On peut généraliser le modèle et les preuves faites ci-dessus au cas général XMDP où l'espace d'état est hybride et les actions paramétriques dépendent de plusieurs paramètres. Dans ce cas, x est le vecteur contenant tous les paramètres. Par exemple, dans le

cas d'un satellite d'observation de la Terre, on peut avoir deux paramètres : un angle θ et une durée τ, le vecteur des paramètres est alors $x = (\theta, \tau)$. Certaines actions, comme "télédécharger les données" ne dépendent pas de x, d'autres, comme "tourner de θ degrés" dépendent d'une partie des éléments du vecteur x, enfin on peut avoir des actions dépendant de plusieurs paramètres comme "tourner à la vitesse $\frac{\theta}{\tau}$ pendant τ secondes". Les politiques pour les XMDP généraux peuvent être optimisées de la même manière qu'à la section précédente (équation 7). Ceci nous permet de représenter une classe plus générale de problèmes de décision, incluant les problèmes MDP standard.

La méthode générale d'optimisation des XMDP basée sur l'itération de la valeur peut être décrite comme l'alternance d'une étape d'optimisation non-linéaire afin de trouver le paramètre optimal par action, suivie d'une étape d'optimisation discrète sur un jeu fini d'actions sachant leur paramètre optimal. Cette méthode utilise la formulation proposée à l'equation 7 afin de construire la séquence des fonctions de valeur $(V_n)_{n\in\mathbb{N}}$:

$$V_{n+1}(s,t) = \max_{a\in A} \Big\{ Q_n(s,t,a) \Big\} \tag{8}$$

$$Q_n(s,t,a) = \sup_{x\in X} \left\{ r(s,t,a(x)) \; + \int_{\substack{t'\in\mathbb{R} \\ s'\in S}} \gamma^{t'-t} p(s',t'|s,t,a(x)) V_n(s',t') ds' dt' \right\} \tag{9}$$

La principale spécificité du modèle à actions paramétriques réside dans l'impact des actions sur les dates de décision : les dates de décision ne sont plus connues à l'avance et on ne peut plus simplement considérer des séquences à pas de temps entier. Malgré cela, le modèle XMDP que l'on propose étend les équations d'optimalité et les méthodes de résolution classiques à la résolution de classes de problèmes plus généraux.

5 Travaux connexes

Une approche similaire de la dépendance au temps et des paramètres d'actions a été étudiée avec le modèle TMDP (Boyan & Littman, 2001). De fait, on peut écrire un TMDP comme un MDP à actions paramétriques particulier, en mentionnant explicitement chaque action là où le modèle TMDP regroupait en fait deux actions à chaque période de décision : une première action d'attente jusqu'à une certaine date et une action a à entreprendre juste après. Une action d'une politique TMDP est spécifiée dans (Boyan & Littman, 2001) comme une paire (t', a) qui signifie que l'action optimale dans l'état courant est "attendre la date t'" puis "entreprendre a". Cela dénote en fait l'existence de deux actions distinctes dans un XMDP, la première étant l'action paramétrique "attendre la date t'" et la seconde étant l'action standard non-paramétrique a. On peut grouper ces deux actions dans le modèle TMDP car l'action "attendre" est considérée déterministe. Malheureusement, (Boyan & Littman, 2001) n'explicite pas le critère utilisé nin'établit d'équations d'optimalité. Notre contribution peut être vue comme une généralisation du modèle TMDP et des MDP classiques à des MDP à actions paramétriques, génériques, avec des dates de décision aléatoires pour lesquels on prouve le lien entre critère γ-pondéré et équation d'optimalité. Le modèle XMDP comble également les lacunes du modèle SMDP+ précédemment présenté dans (Rachelson *et al.*, 2006) où la spécification de l'action attendre n'était pas précise. De ce fait, les XMDP peuvent

être vus comme un cadre pour modéliser et résoudre les problèmes MDP instationnaires sans distincition entre horizon fini et infini.

Il est également intéressant de noter le lien avec les travaux sur les Generalized SMDP (Younes & Simmons, 2004). Les GSMDP consituent un modèle efficace pour gérer les actions et les évènements à durée aléatoire et concurrents. On pourrait tenter d'écrire un MMDP ou un CoMDP paramétrique sur le modèle de (Boutilier, 1996; Mausam & Weld, 2005) afin de prendre en compte des actions concurrentes. Toutefois, la prise en compte d'évènements, comme dans les GSMDP, est un sujet qui sort de ce cadre.

Le fait de considérer des actions à paramètres continus nous a amenés à prendre en compte une variable temporelle continue et à définir des effets continus sur l'espace d'état. On pourrait, pour se rapprocher du cadre classique MDP, discrétiser dans une certaine mesure cet espace d'état mais la discrétisation des étapes de décision ramène les XMDP à des MDP standards de très grande taille du fait de la discrétisation. Par ailleurs, de nombreux travaux ont été entrepris afin de résoudre les MDP à espace d'états hybride par programmation linéaire approchée (ALP) (Hauskrecht & Kveton, 2006; Guestrin *et al.*, 2004). L'extension de l'algorithme HALP au cadre des actions paramétriques est une voie de recherche intéressante.

Des exemples de cas d'application des XMDP peuvent être trouvés dans (Hanks *et al.*, 1995) avec l'illustration de l'agent d'aide à la décision médicale devant proposer une certaine dose d'une certaine substance à injecter à un blessé. D'autres exemples, comme mentionné précédemment, existent dans le cadre des applications satellitaires, en particulier pour les satellites agiles capables de pointer dans des directions différentes depuis un point de leur orbite (Beaumet, 2006). Pour ces satellites, on doit décider de la durée des observations, de l'amplitude des rotations, etc., en même temps que l'on décide effectivement d'observer, de tourner ou autre : on doit donc décider à la fois de l'action à entreprendre et de la valeur de son paramètre, faisant de ce problème un cas de test intéressant pour les XMDP. Enfin, l'exemple du rover présenté en introduction, similaire au problème de (Boyan & Littman, 2001), consitue le principal cas auquel on s'intéresse actuellement. L'implantation d'algorithmes d'optimisation exacte et approchée de stratégies dépendantes du temps est en cours.

6 Conclusion

Afin de traiter les problèmes où choisir les paramètres d'une action peut être critique et aussi important que de choisir l'action elle-même, on a introduit le cadre des MDP à actions paramétriques (XMDP). Ce modèle permet de représenter des problèmes de décision dans l'incertain où l'ajustement des paramètres d'action au problème peut s'avérer nécéssaire pour générer des stratégies efficaces, en particulier dans le cadre des problèmes instationnaires. Nous nous sommes attachés à montrer que les équations générales d'optimalité étaient toujours valables dans le cadre XMDP et nous les avons étendues du cadre des MDP classiques au modèle à actions paramétriques et espace d'états hybride. Sous les hypothèses mentionnées en section 2, nous avons montré ainsi que les résultats MDP se prolongeaient au cadre XMDP, en particulier l'équivalence

entre critère γ-pondéré et l'équation d'optimalité de Bellman généralisée (equation 6).

La principale spécificité du modèle à actions paramétriques réside dans la possibilité de considérer des dates de décision aléatoires en prenant en compte un temps explicite et observable. Cet aspect nous permet de planifier dans l'incertain dans des domaines instationnaires et de considérer un critère γ-pondéré cohérent avec des séquences d'actions de durées aléatoires. Outre le traitement du paramètre de durée, le modèle XMDP permet de traiter des problèmes plus généraux à actions paramétriques continues comme ceux présentés en section 5.

Les travaux actuels se concentrent sur l'implantation et le test du modèle XMDP dans le cas de problèmes dépendants du temps. Nous avons développé un cadre basé sur les approximations par des polynômes définis par morceaux des fonctions de valeurs de politiques générées pour des problèmes inspirés de (Boyan & Littman, 2001). Une autre voie de recherche que nous souhaitons explorer implique l'utilisation du modèle des actions paramétriques dans le cadre de stratégies dépendantes du temps afin de s'intéresser à la coordination d'actions concurrentes.

Références

BEAUMET G. (2006). Continuous planning for the control of an autonomous agile satellite. In *International Conference on Automated Planning and Scheduling*.

BELLMAN R. (1957). *Dynamic Programming*. Princeton University Press, New Jersey.

BERTSEKAS D. P. & SHREVE S. E. (1996). *Stochastic Optimal Control : The Discrete-Time Case*. Athena Scientific.

BOUTILIER C. (1996). Planning, learning and coordination in multiagent decision processes. In *Theoretical Aspects of Rationality and Knowledge*, p. 195–201.

BOYAN J. A. & LITTMAN M. L. (2001). Exact solutions to time dependent MDPs. *Advances in Neural Information Processing Systems*, **13**, 1026–1032.

BRESINA J., DEARDEN R., MEULEAU N., RAMAKRISHNAN S. & WASHINGTON R. (2002). Planning under continuous time and resource uncertainty : a challenge for AI. In *19th Conf. on Uncertainty in AI*.

FENG Z., DEARDEN R., MEULEAU N. & WASHINGTON R. (2004). Dynamic programming for structured continuous Markov decision problems. In *20th Conference on Uncertainty in AI*.

GUESTRIN C., HAUSKRECHT M. & KVETON B. (2004). Solving factored MDPs with continuous and discrete variables. In *20th Conference on Uncertainty in Artificial Intelligence*.

HANKS S., MADIGAN D. & GAVRIN J. (1995). Probabilistic temporal reasoning with endogeneous change. In *11th Conf. on Uncertainty in Artificial Intelligence*.

HAUSKRECHT M. & KVETON B. (2006). Approximate linear programming for solving hybrid factored MDPs. In *9th International Symposium on Artificial Intelligence and Mathematics*.

MAUSAM & WELD D. (2005). Concurrent probabilistic temporal planning. In *International Conference on Automated Planning and Scheduling*.

PUTERMAN M. L. (1994). *Markov Decision Processes*. John Wiley & Sons, Inc.

RACHELSON E., FABIANI P., FARGES J., TEICHTEIL F. & GARCIA F. (2006). Une approche du traitement du temps dans le cadre MDP : trois méthodes de découpage de la droite temporelle. In *Journées Françaises Planification Décision Apprentissage*.

YOUNES H. L. S. & SIMMONS R. G. (2004). Solving generalized semi-Markov processes using continuous phase-type distributions. In *19th National Conf. on Artificial Intelligence*.