Planification dans l'incertain
—
Introduire une variable temporelle continue

Emmanuel RACHELSON

ONERA/DCSD

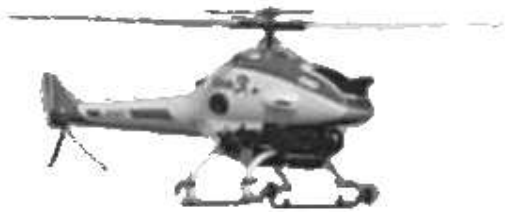
11 mai 2006

Plan

Plan

Introduction de la problématique

Mission d'observation



Introduction de la problématique

Mission d'observation

- Incertitude sur résultats actions : modèle probabiliste
  $\longrightarrow$ solution robuste sous forme de politiques.

Introduction de la problématique

Mission d'observation

- Incertitude sur résultats actions : modèle probabiliste
  $\longrightarrow$ solution robuste sous forme de politiques.
- Instationnarité du modèle $\longrightarrow$ politiques instationnaires.

Introduction de la problématique

Mission d'observation

- Incertitude sur résultats actions : modèle probabiliste
  $\longrightarrow$ solution robuste sous forme de politiques.
- Instationnarité du modèle $\longrightarrow$ politiques instationnaires.
- Bonne stratégie ? Inuitivement : " jusqu'à t_1 (à déterminer), aller chercher le passager, après, ne rien faire".
  - $\longrightarrow$ Dépendance explicite de la stratégie au temps
  - $\longrightarrow$ Compacité de la solution malgré l'introduction d'un paramètre supplémentaire.

Introduction de la problématique

Généralisation

0n cherche à résoudre des problèmes dont le modèle d'univers est instationnaire, stationnaire à l'infini, c'est-à-dire :

- Les $\begin{cases} \text{changements d'état} \\ \text{durées de transition} \\ \text{récompenses disponibles} \end{cases}$ dépendent de la date courante.
- $\exists T \in \mathbb{R}^+ / \forall t > T$ "problème(t)=problème(T)"

Note : On appelle le plus petit des T "pseudo-horizon".

Plan

Le modèle MDP

Incertitude sur résultat des actions $\longrightarrow$ modèle probabiliste de l'univers

Le modèle MDP

Incertitude sur résultat des actions $\longrightarrow$ modèle probabiliste de l'univers

- Espace d'état S
- Espace d'action A
- Modèle de transition $P(s'|s,a)$
- Modèle de récompense $r(s,a)$

Le modèle MDP

Incertitude sur résultat des actions $\longrightarrow$ modèle probabiliste de l'univers

- Espace d'état S
- Espace d'action A
- Modèle de transition $P(s'|s,a)$
- Modèle de récompense $r(s,a)$

Solutions : politiques $\pi : S \longrightarrow A$: actions à entreprendre

Valeur d'une politique : $V^{\pi} : S \longrightarrow \mathbb{R}$: récompense espérée

Le modèle MDP

Incertitude sur résultat des actions $\longrightarrow$ modèle probabiliste de l'univers

- Espace d'état S
- Espace d'action A
- Modèle de transition $P(s'|s,a)$
- Modèle de récompense $r(s,a)$

Solutions : politiques $\pi : S \longrightarrow A$: actions à entreprendre
Valeur d'une politique : $V^\pi : S \longrightarrow \mathbb{R}$: récompense espérée

Critère γ-pondéré, équation de Bellman :

$$V^\pi(s) = E\left(\sum_{t=0}^{\infty} \gamma^t r_t^\pi | s\right)$$

$$\forall s \in S, V^*(s) = \max_{a \in A}\left\{r(s,a) + \gamma \cdot \sum_{s' \in S} P(s'|s,a)V^*(s')\right\}$$

Le modèle SMDP

On intègre la notion de durée d'action.

Le modèle SMDP

On intègre la notion de durée d'action.

- Espaces d'état et d'action : S, A
- Modèle de transition : $Q(t, j|s, a)$
- Modèle de récompense $k(s, a)$ (instantané) et $c(j, s, a)$ (taux de récompense)

Le modèle SMDP

On intègre la notion de durée d'action.

- Espaces d'état et d'action : S, A
- Modèle de transition : $Q(t, j|s, a)$
- Modèle de récompense $k(s, a)$ (instantané) et $c(j, s, a)$ (taux de récompense)

$$V^* = \max_{\pi} \left\{ r_\pi(s) + \sum_{s' \in S} m_\pi(s', s) V^*(s') \right\}, m_\pi(s', s) = \int_0^\infty \gamma^t Q_\pi(dt, s'|s)$$

Le modèle SMDP

On intègre la notion de durée d'action.

- Espaces d'état et d'action : S, A
- Modèle de transition : $Q(t, j|s, a)$
- Modèle de récompense $k(s, a)$ (instantané) et $c(j, s, a)$ (taux de récompense)

$$V^* = \max_{\pi} \left\{ r_\pi(s) + \sum_{s' \in S} m_\pi(s', s) V^*(s') \right\}, m_\pi(s', s) = \int_0^\infty \gamma^t Q_\pi(dt, s'|s)$$

Pb : Le pb est toujours stationnaire, on souhaite pouvoir spécifier des politiques du type $\pi(s, t)$

$\longrightarrow$ Pour cela, il faut rendre la variable temporelle observable par l'agent *ie.* l'inclure dans son espace d'état.

Un modèle avec dates

On ajoute une variable temporelle continue et observable

Un modèle avec dates

On ajoute une variable temporelle continue et observable

- Etat augmenté $\sigma = (s, t)$
- Actions $a \in A$
- Transitions : $P(\sigma'|\sigma, a) = P(s'|\sigma, a) \cdot F(t'|\sigma, a)$
- Récompenses : $r(\sigma'|\sigma, a)$

Un modèle avec dates

On ajoute une variable temporelle continue et observable

- Etat augmenté $\sigma = (s, t)$
- Actions $a \in A$
- Transitions : $P(\sigma'|\sigma, a) = P(s'|\sigma, a) \cdot F(t'|\sigma, a)$
- Récompenses : $r(\sigma'|\sigma, a)$

Conventions de notation : $\begin{cases} t,\, t' \rightarrow \text{dates} \\ \tau \rightarrow \text{durées} \end{cases}$

Un modèle avec dates

On ajoute une variable temporelle continue et observable

- Etat augmenté $\sigma = (s,t)$
- Actions $a \in A$
- Transitions : $P(\sigma'|\sigma,a) = P(s'|\sigma,a) \cdot F(t'|\sigma,a)$
- Récompenses : $r(\sigma'|\sigma,a)$

Conventions de notation : $\begin{cases} t,\, t' \rightarrow \text{dates} \\ \tau \rightarrow \text{durées} \end{cases}$

L'équation de Bellman devient :

$$V^*(\sigma) = \max_{a \in A} \left\{ \sum_{s' \in S} \int_0^\infty [r(s', t+\tau|\sigma,a) + \gamma^\tau V^*(\sigma')] F(\tau|\sigma,a) P(s'|\sigma,a) d\tau \right\}$$

Comment résoudre un tel problème ?

Plan

Résolution par discrétisation

F distribution discrète sur les durées

Note : on peut transformer un F continu en F discret.

Résolution par discrétisation

F distribution discrète sur les durées

Note : on peut transformer un F continu en F discret.

Soit θ le *pgcd* des durées d'action. On ajoute une variable d'état t prenant les valeurs : $\{0, \theta, \ldots, (n+1)\theta\}$ avec $n = floor(\frac{T}{\theta})$.

Résolution par discrétisation

F distribution discrète sur les durées

Note : on peut transformer un F continu en F discret.

Soit θ le *pgcd* des durées d'action. On ajoute une variable d'état t prenant les valeurs : $\{0, \theta, \ldots, (n+1)\theta\}$ avec $n = floor(\frac{T}{\theta})$.

On redéfinit le modèle de récompense et de transition sur $S \cup \{t\}$ et on se ramène au cas discret.

$\rightarrow$ intérêt : les matrices de transition sont creuses.

Résolution par discrétisation

F distribution discrète sur les durées

Note : on peut transformer un F continu en F discret.

Soit θ le *pgcd* des durées d'action. On ajoute une variable d'état t prenant les valeurs : $\{0, \theta, \ldots, (n+1)\theta\}$ avec $n = floor(\frac{T}{\theta})$.

On redéfinit le modèle de récompense et de transition sur $S \cup \{t\}$ et on se ramène au cas discret.

$\longrightarrow$ intérêt : les matrices de transition sont creuses.

On obtient bien une politique en $\pi(s,t)$ comme souhaité mais ...

- ... la compacité n'y est pas.
- ... dès que le problème devient complexe et T grand, le cardinal de t explose.
- ... on pert en optimalité lors de la discrétisation de t.

Plan

Hypothèses

Famille de fonctions exp-poly :

$$f \text{ est exp-poly} \Leftrightarrow \exists(n,\alpha) \in \mathbb{N} \times \mathbb{R} / \forall x \in \mathbb{R} \quad f(x) = e^{\alpha x} \cdot P(x), P \in \mathscr{P}_n.$$

Hypothèses

Famille de fonctions exp-poly :

$$f \text{ est exp-poly} \Leftrightarrow \exists(n,\alpha) \in \mathbb{N} \times \mathbb{R} / \forall x \in \mathbb{R} \quad f(x) = e^{\alpha x} \cdot P(x), P \in \mathscr{P}_n.$$

3 hypothèses :

H1 Disponibilité des récompenses
$r(\sigma', a, \sigma) = \sum_{j \in S} 1_j(s') \cdot rr_j(s,a) \cdot rt_j(t')$, avec rt_j exp-poly,
eventuellement discontinue.

Hypothèses

Famille de fonctions exp-poly :

$$f \text{ est exp-poly} \Leftrightarrow \exists(n,\alpha) \in \mathbb{N} \times \mathbb{R} / \forall x \in \mathbb{R} \quad f(x) = e^{\alpha x} \cdot P(x), P \in \mathscr{P}_n.$$

3 hypothèses :

H1 Disponibilité des récompenses
$r(\sigma', a, \sigma) = \sum_{j \in S} 1_j(s') \cdot rr_j(s,a) \cdot rt_j(t')$, avec rt_j exp-poly,
eventuellement discontinue.

H2 Stationnarité des durées et probabilités
$\forall t \in \mathbb{R}^+, F(t'|\sigma,a) = F(\tau|s,a)$ et $P(s'|\sigma,a) = P(s'|s,a)$.

Hypothèses

Famille de fonctions exp-poly :

$$f \text{ est exp-poly} \Leftrightarrow \exists (n,\alpha) \in \mathbb{N} \times \mathbb{R} / \forall x \in \mathbb{R} \quad f(x) = e^{\alpha x} \cdot P(x), P \in \mathscr{P}_n.$$

3 hypothèses :

H1 Disponibilité des récompenses
$r(\sigma', a, \sigma) = \sum_{j \in S} 1_j(s') \cdot rr_j(s,a) \cdot rt_j(t')$, avec rt_j exp-poly,
eventuellement discontinue.

H2 Stationnarité des durées et probabilités
$\forall t \in \mathbb{R}^+, F(t'|\sigma,a) = F(\tau|s,a)$ et $P(s'|\sigma,a) = P(s'|s,a)$.

H3 $F(\tau|s,a)$ s'écrit comme :

cas 1 : Une fonction constante par morceaux

cas 2 : Une distribution à densité exp-poly

Le résultat utile

Propriété

La primitive d'une fonction exp-poly est exp-poly de même coefficient exponentiel et de même degré.

$$\Rightarrow V^{\pi} \text{ est exp-poly de degré } \max_j \left\{ deg(rt_j \cdot F) \right\}$$

car, selon les hypothèses précédentes :

$$V^{\pi}(\sigma) = \sum_{s' \in S} \int_0^{\infty} \left(rr_{s'}(s, \pi(s)) \cdot rt_{s'}(t') + \gamma^{\tau} V^{\pi}(\sigma) \right) \cdot F(\tau | s, \pi(s)) P(s' | s, \pi(s))) \, d\tau$$

On en tire l'algorithme suivant.

L'algorithme

pour $s \in S$ **faire**
    $Lpi(s) \leftarrow \{(0,[0,\ldots,0])\}$
fin
répéter
    pour $s \in S$ **faire**
        pour $a \in A$ **faire**
            $La(s,a) \leftarrow calcul(Lpi(s),r,F,P)$
        fin
    fin
    pour $s \in S$ **faire**
        $Lpi(s) \leftarrow La(s,a_0)$
        pour $a \in A$ **faire**
            $Lpi(s) \leftarrow enveloppe(Lpi(s),La(s,a))$
        fin
    fin
jusqu'à *pseudo-optimalité*
retourner *Lpi(s)*

faiblesses de l'algorithme exp-poly

L'ordre de grandeur du nombre de coefficients conservés en cache :

$$\left(deg\left(rt\left(t\right)\right)+2\right)\cdot\left|S\right|^2\cdot\left|A\right|\cdot n$$

(n : nombre moyen de morceaux de $rt_j \cdot F$)

calcul de $\left|A\right|$ intersections de fonctions exp-poly

$\longrightarrow$ racines de $n\cdot\left|A\right|$ polynômes de degré $deg\left(rt\right)$.

Plan

L'idée

Rechercher les dates de décision "importantes"

"importantes" = la décision optimale change en cette date.

L'idée

Rechercher les dates de décision "importantes"
"importantes" = la décision optimale change en cette date.

Objectif : manipuler des objets aussi simples (compacts) que la solution que l'on souhaite.

L'idée

Rechercher les dates de décision "importantes"
"importantes" = la décision optimale change en cette date.

Objectif : manipuler des objets aussi simples (compacts) que la solution que l'on souhaite.

Idée : Réutiliser les résultats et modèles des sections précédentes,
Construire processus itératif,
Chercher les dates de décision optimales
et les décisions associées, par état.

Définitions

Définition (Période de décision)

Intervalle temporel sur lequel, pour un état donné, l'action à entreprendre est constante.

Définition (Période de décision optimale)

Une période de décision est optimale si l'action à y entreprendre est toujours l'action optimale.

Définitions

Définition (Ensemble $\tilde{T}$)

On assimile les dates de début de période de décision et la période elle-même et on note $\tilde{T}$ l'ensemble de ces dates.

Définition (Ensemble $\tilde{T}_s$)

On note $\tilde{T}_s$ l'ensemble des dates de décision utiles en un état s.

Note : Une période de décision donnée par sa date de début est toujours relative à un état !

Définitions

Définition (Politique)

Une politique est une application de $S \times \{T_s\}$ dans A.

Définition (t-erreur de Bellman)

On définit la t-erreur de Bellman en s comme :

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

Note : La t-erreur de Bellman fournit, en une date t, la valeur de combien on peut améliorer la politique courante en optimisant sur un coup.

Initialisation

Initiallement :

$$\forall s \in S, \tilde{T} = \tilde{T}_s = \{0, T\}$$

$\longrightarrow$ deux périodes de décision, avant et après le pseudo-horizon.

4 étapes

Etape 1 : Génération de $\tilde{P}$ et $\tilde{r}$.

$P, F, r, \{\tilde{T}_s, s \in S\} \quad \rightarrow \quad \tilde{P}, \tilde{r}$

Nouveau MDP discret $\tilde{M} = \left\{S \times \{\tilde{T}_s, s \in S\}, A, \tilde{P}, \tilde{r}\right\}$.

4 étapes

Etape 1 : Génération de $\tilde{P}$ et $\tilde{r}$.

$P, F, r, \{\tilde{T}_s, s \in S\} \quad \rightarrow \quad \tilde{P}, \tilde{r}$

Nouveau MDP discret $\tilde{M} = \left\{S \times \{\tilde{T}_s, s \in S\}, A, \tilde{P}, \tilde{r}\right\}$.

Etape 2 : Optimisation de $\tilde{M}$.

Pour un $s \in S$, on génère $\left|\tilde{T}_s\right|$ actions.

$\pi(s, \tilde{t})$, $V^{\pi}(s, \tilde{t})$

4 étapes

Etape 1 : Génération de $\tilde{P}$ et $\tilde{r}$.

$P, F, r, \{\tilde{T}_s, s \in S\} \quad \rightarrow \quad \tilde{P}, \tilde{r}$

Nouveau MDP discret $\tilde{M} = \left\{S \times \{\tilde{T}_s, s \in S\}, A, \tilde{P}, \tilde{r}\right\}$.

Etape 2 : Optimisation de $\tilde{M}$.

Pour un $s \in S$, on génère $\left|\tilde{T}_s\right|$ actions.

$\pi(s, \tilde{t})$, $V^{\pi}(s, \tilde{t})$

Etape 3 : Recherche des dates de décision optimales.

$\pi(s, \tilde{t}) \quad \rightarrow \quad \pi(s, t) \quad \rightarrow \quad V^{\pi}(s, t)$

$\underset{t \in [0,T]}{max}\ BE_s(t) \quad \rightarrow \quad (s, t_s, \varepsilon_s)$

4 étapes

Etape 1 : Génération de $\tilde{P}$ et $\tilde{r}$.

$P, F, r, \{\tilde{T}_s, s \in S\} \quad \rightarrow \quad \tilde{P}, \tilde{r}$

Nouveau MDP discret $\tilde{M} = \left\{S \times \{\tilde{T}_s, s \in S\}, A, \tilde{P}, \tilde{r}\right\}$.

Etape 2 : Optimisation de $\tilde{M}$.

Pour un $s \in S$, on génère $\left|\tilde{T}_s\right|$ actions.

$\pi(s, \tilde{t})$, $V^{\pi}(s, \tilde{t})$

Etape 3 : Recherche des dates de décision optimales.

$\pi(s, \tilde{t}) \quad \rightarrow \quad \pi(s, t) \quad \rightarrow \quad V^{\pi}(s, t)$

$\underset{t \in [0,T]}{max}\ BE_s(t) \quad \rightarrow \quad (s, t_s, \varepsilon_s)$

Etape 4 : Peuplement des $\tilde{T}_s$

Si $\varepsilon_s > \varepsilon$ alors $\tilde{T}_s \leftarrow \tilde{T}_s \cup \{t_s\}$.

Itérations

Entre étape 2 et 3 : simplification des $\tilde{T}_s$

$\longrightarrow$ Cache de dates de décision minimal

Itérations

Entre étape 2 et 3 : simplification des $\tilde{T}_s$

$\longrightarrow$ Cache de dates de décision minimal

Condition d'arrêt :

- $\forall s \in S, \varepsilon_s < \varepsilon$

  Note : il n'est pas garanti que cette condition soit jamais remplie.
- Arrêt par l'utilisateur (fonctionnement "anytime")

Optimisation de l'erreur de Bellman

Point crucial de l'algorithme

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

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

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

Optimisation de l'erreur de Bellman

Deux solutions :

- Résolution numérique
  Avantage : ne dépend pas du modèle. Défaut : pas de garantie d'optimalité (résolution)
- Modèle exp-poly
  On sait calculer $L_a^t(V^\pi)(s,t)$ simplement car :
  $\begin{cases} \text{plus d'écueil du calcul de l'enveloppe} \\ \text{fonctions exp-poly} \end{cases}$

  $|A|$ maxima des $L_a^t(V^\pi)(s,t) - V^\pi(s,t) \longrightarrow (s, t_s, \varepsilon_s)$

Adaptation à un modèle discret

Hypothèses simplificatrices possibles :

P, F, r définies de façon discrète

F déterministe

F et P stationnaires (cas exp-poly)

Plan

Conclusion

- Définition des problèmes : extension de ppddl et pddl 2.1.
- Implémentation en cours, premiers essais “à la main" encourageants
- Objectif dans un cadre plus large : réintégrer cet algorithme dans la boucle de décision bi-(multi-)agents afin de faciliter la planification en décentralisé et distribué.

Conclusion

- Définition des problèmes : extension de ppddl et pddl 2.1.
- Implémentation en cours, premiers essais “à la main" encourageants
- Objectif dans un cadre plus large : réintégrer cet algorithme dans la boucle de décision bi-(multi-)agents afin de faciliter la planification en décentralisé et distribué.

Conclusion

- Définition des problèmes : extension de ppddl et pddl 2.1.
- Implémentation en cours, premiers essais “à la main" encourageants
- Objectif dans un cadre plus large : réintégrer cet algorithme dans la boucle de décision bi-(multi-)agents afin de faciliter la planification en décentralisé et distribué.