Planifier en fonction du temps dans le cadre MDP

Emmanuel Rachelson [1]

Frederick Garcia [2]
Florent Teichteil [1]

[1]ONERA-DCSD — Toulouse

[2]INRA-BIA —Toulouse

25 mai 2007

Plan

Problématique

- Exemples
- Représenter l'incertain et l'instationnaire

Modèles

- Modèles classiques
- SMDP, SMDP+ et TMDP
- XMDP, un modèle générique

Optimisation des politiques

- TMDPpoly - DP
- SMDP+ - BellErr

Conclusion

Plan

Problématique
- Exemples
- Représenter l'incertain et l'instationnaire

Modèles
- Modèles classiques
- SMDP, SMDP+ et TMDP
- XMDP, un modèle générique

Optimisation des politiques
- TMDPpoly - DP
- SMDP+ - BellErr

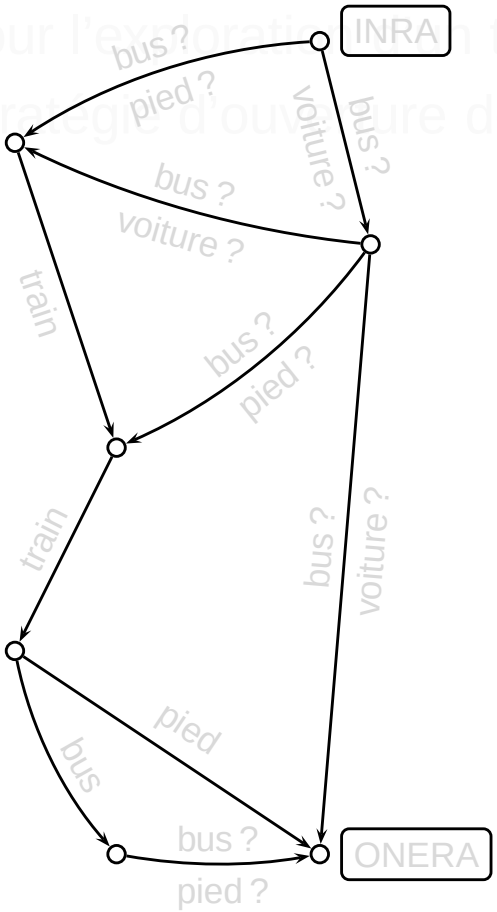
Conclusion

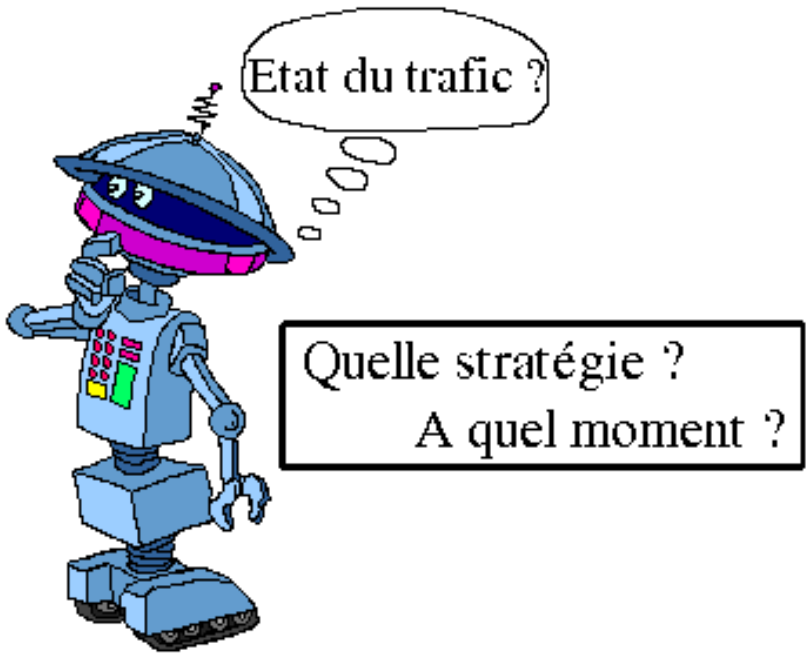
Problèmes types

- Gestion des activités d'un satellite
- Optimisation de la statégie de déplacement ONERA → INRA
- Coordination des actions d'un binôme d'agents aéroterrestre pour l'exploration d'un terrain
- Stratégie d'ouverture des guichets d'un aéroport

Problèmes types

- Gestion des activités d'un satellite
- Optimisation de la statégie de déplacement ONERA → INRA
- Coordination des actions d'un binôme d'agents aéroterrestre pour l'exploration d'un terrain
- Stratégie d'ouverture des guichets d'un aéroport





Problèmes types

- Gestion des activités d'un satellite
- Optimisation de la statégie de déplacement ONERA $\longrightarrow$ INRA
- Coordination des actions d'un binôme d'agents aéroterrestre pour l'exploration d'un terrain
- Stratégie d'ouverture des guichets d'un aéroport

Problèmes types

- Gestion des activités d'un satellite
- Optimisation de la statégie de déplacement ONERA $\longrightarrow$ INRA
- Coordination des actions d'un binôme d'agents aéroterrestre pour l'exploration d'un terrain
- Stratégie d'ouverture des guichets d'un aéroport

Plan

Problématique
- Exemples
- Représenter l'incertain et l'instationnaire

Modèles
- Modèles classiques
- SMDP, SMDP+ et TMDP
- XMDP, un modèle générique

Optimisation des politiques
- TMDPpoly - DP
- SMDP+ - BellErr

Conclusion

Représenter l'incertain et l'instationnaire

incertain $\longrightarrow$ Processus Décisionnels de Markov

instationnaire $\longrightarrow$ quelles options de modélisation et de résolution ?

Temps discret / continu

Definition (date de décision)

Instant du processus où une décision est prise. Les transitions du processus amènent de date de décision en date de décision.

Temps discret / continu

Definition (date de décision)

Instant du processus où une décision est prise. Les transitions du processus amènent de date de décision en date de décision.

Décider en fcn du temps $\leftrightarrow$ temps observable $\leftrightarrow$ t, variable d'état

Temps discret / continu

Definition (date de décision)

Instant du processus où une décision est prise. Les transitions du processus amènent de date de décision en date de décision.

Décider en fcn du temps $\leftrightarrow$ temps observable $\leftrightarrow$ t, variable d'état

MDP classique : dates de décision entières et déterministes,

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

Temps discret / continu

Definition (date de décision)

Instant du processus où une décision est prise. Les transitions du processus amènent de date de décision en date de décision.

Décider en fcn du temps $\leftrightarrow$ temps observable $\leftrightarrow$ t, variable d'état

MDP classique : dates de décision entières et déterministes,

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

Durées discrètes + incertaines $\Rightarrow$ explosion de l'espace d'état

Temps discret / continu

Definition (date de décision)

Instant du processus où une décision est prise. Les transitions du processus amènent de date de décision en date de décision.

Décider en fcn du temps $\leftrightarrow$ temps observable $\leftrightarrow$ t, variable d'état

MDP classique : dates de décision entières et déterministes,

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

Durées discrètes + incertaines $\Rightarrow$ explosion de l'espace d'état

Temps continu + durées incertaines $\Rightarrow$ décider des *périodes de décision*, problème de découpage de la droite temporelle.

ONERA THE FRENCH AEROSPACE LAB INRA

Période de décision

Definition (Période de décision)

Une période de décision pour une politique d'un MDP à temps continu est un intervalle sur lequel l'action à entreprendre définie par la politique est constante.

Trouver les périodes de décision optimales selon un certain critère, c'est décider *quoi* faire et *quand*.

Horizon fini / infini

MDP classique : horizon fini $\Leftrightarrow$ nombre de 'coups' fini.
Nombre de coups ou temps d'exécution fini ?

On va s'intéresser à des processus à nombre d'actions non limité.

$\rightarrow$ on couvre les cas à temps d'exécution fini (planifier en fonction d'une ressource "temps restant")

Dès qu'on parle d'instationnarité, il y a :

- ou bien un horizon où s'arrête l'instationnarité (*pseudo-horizon*)
- ou bien une notion de périodicité

Enjeu du problème

Ecrire un problème “type MDP”
avec une variable d’état “temps” continue et observable
ce qui permet de considérer le problème comme stationnaire (le modèle du problème ne change pas entre l’action *i* et l’action *j*)

Difficulté : trouver les périodes de décision.

Plan

Problématique
- Exemples
- Représenter l'incertain et l'instationnaire

Modèles
- Modèles classiques
- SMDP, SMDP+ et TMDP
- XMDP, un modèle générique

Optimisation des politiques
- TMDPpoly - DP
- SMDP+ - BellErr

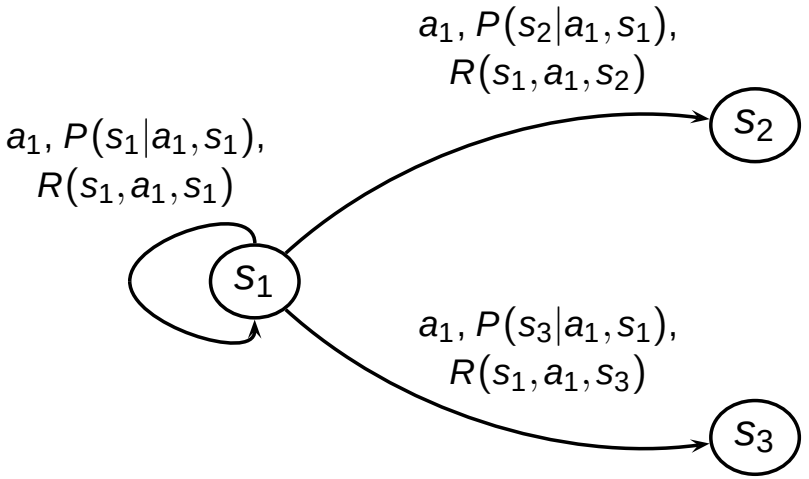
Conclusion

Modèles existants

- STDN : incert. durée, temps continu, stat.
- SMDP : incert. durée, incert. action, temps continu, stat. (Puterman, 94)
- TMDP : incert. durée, incert. action, temps continu, instat. (Boyan and Littman, 01)

Rappels MDP

$$\text{MDP} : < S, A, P, r, T >$$



$$\pi(s) \leftrightarrow V^{\pi}(s)$$

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

Plan

Problématique
- Exemples
- Représenter l'incertain et l'instationnaire

Modèles
- Modèles classiques
- SMDP, SMDP+ et TMDP
- XMDP, un modèle générique

Optimisation des politiques
- TMDPpoly - DP
- SMDP+ - BellErr

Conclusion

SMDP et SMDP+

SMDP : $< S, A, P, F, r, T >$

- S espace d'états
- A espace d'actions
- Modèle de transition $Q(s', t'|s, a) = P(s'|s, a) \cdot F(t'|s, a)$
- Modèle de récompense $r(s, a)$

$\Rightarrow$ découplage $s' \leftrightarrow t'$ et indépendance en t

---

SMDP+ : on retire ces hypothèses :

- état $\sigma = (s, t) \in \Sigma = S \times \mathbb{R}$ (état discret + temps continu)
- A espace d'actions
- $Q(\sigma'|\sigma, a) = Q(s', t'|s, t, a) = P(s'|s, t, a) \cdot F(t'|s, t, a, s')$
- Modèle de récompense $r(s, t, a)$

politique SMDP+

Definition (politique SMDP+)

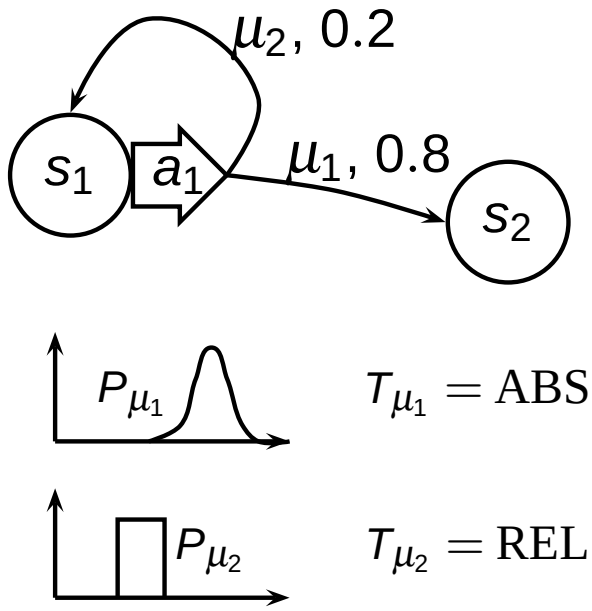
Application de $\Sigma = S \times \mathbb{R}$ dans A qui, à chaque couple état discret - date courante, associe une action à entreprendre.

On suppose que pour un problème réel il existe un nombre fini de périodes de décision, on cherche donc le découpage optimal de la droite temporelle en même temps que l'action optimale à entreprendre sur chaque période de décision.

Problème : pas d'action "attendre" (solution : XMDP).

TMDP - Boyan and Littman, 01

- S, A
- $L(\mu|s,t,a)$
- $P_\mu(t'), P_\mu(t'-t)$
- $r(\mu,t,t')$
- $K(s,\theta)$



- politique : $(s,t) \mapsto (t',a)$
- En pratique, hypothèses restrictives :
  - P_μ densité d'une distribution discrète
  - $L(\mu|s,t,a)$ constante par morceaux
  - $r(\mu,t), r(\mu,t'), r(\mu,t'-t)$ linéaires par morceaux.

TMDPpoly

Extension de TMDP à des fonctions P_μ, L et r polynômiales par morceaux.

Intérêt pour la résolution et l'expressivité du modèle

Problèmes : pas d'action "attendre" définie dans le modèle (définie dans la méthode de résolution, comme SMDP+) et critère total.

Lien SMDP+ $\leftrightarrow$ TMDP ?

$$M(s,a) = \{\mu \in M \ / \exists t \in \mathbb{R} \text{ tel que } L(\mu|s,t,a) \neq 0\}$$

$$M(s',s,a) = \left\{\mu \in M(s,a) \ / s'_\mu = s'\right\}$$

$$r(s',t',a,s,t) = \sum_{\mu \in M(s',s,a)} r(\mu,t',t)$$

$$P(s'|s,t,a) = \sum_{\mu \in M(s',s,a)} L(\mu|s,t,a)$$

$$F(t'|s,t,a,s') = \sum_{\mu \in M(s',s,a)} P_\mu(t',t)$$

Plan

Problématique
- Exemples
- Représenter l'incertain et l'instationnaire

Modèles
- Modèles classiques
- SMDP, SMDP+ et TMDP
- XMDP, un modèle générique

Optimisation des politiques
- TMDPpoly - DP
- SMDP+ - BellErr

Conclusion

Considérations générales

- ↪ durées continues + probas $\Rightarrow$ effet continu $\Rightarrow$ état continu.
  ex : effet sur la variable d'état "position" de l'action "conduire pendant τ heures" à la vitesse V. Ou plus simplement : "attendre τ secondes".
- ↪ notion d'action paramétrique
- ↪ τ : paramètre d'action spécial vis-à-vis du critère γ-pondéré.

Définition d'un XMDP

MDP à actions paramétriques ou XMDP : $S, A(X), p, r, T$

- Espace d'états hybride factorisé S incluant la variable temporelle.
- Espace d'actions paramétriques $a(x) \in A(X)$ dénombrable.
- Modèle de transition $p(s'|s, a(x))$.
- Modèle de récompense $r(s, a(x))$.

Définition d'un XMDP

MDP à actions paramétriques ou XMDP : $S, A(X), p, r, T$

- Espace d'états hybride factorisé S incluant la variable temporelle.
- Espace d'actions paramétriques $a(x) \in A(X)$ dénombrable.
- Modèle de transition $p(s'|s, a(x))$.
- Modèle de récompense $r(s, a(x))$.

Définition d'un XMDP

MDP à actions paramétriques ou XMDP : $S, A(X), p, r, T$

- Espace d'états hybride factorisé S incluant la variable temporelle.
- Espace d'actions paramétriques $a(x) \in A(X)$ dénombrable.
- Modèle de transition $p(s'|s, a(x))$.
- Modèle de récompense $r(s, a(x))$.

Définition d'un XMDP

MDP à actions paramétriques ou XMDP : $S, A(X), p, r, T$

- Espace d'états hybride factorisé S incluant la variable temporelle.
- Espace d'actions paramétriques $a(x) \in A(X)$ dénombrable.
- Modèle de transition $p(s'|s, a(x))$.
- Modèle de récompense $r(s, a(x))$.

Equation d'optimalité

Hypothèses :

- r **est bornée,**
- r **est semi-continue supérieure,**
- **les durées de transition sont minorées par un réel strictement positif.**

Critère γ-pondéré :

$$V^{\pi}(s) = E\left(\sum_{i=1}^{\infty} \gamma^{t_i} r(s_i, t_i, \pi(s_i, t_i))\right)$$

Caractérisation des politiques optimales :

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

Equation d'optimalité

Hypothèses :

- r **est bornée,**
- r **est semi-continue supérieure,**
- **les durées de transition sont minorées par un réel strictement positif.**

Critère γ-pondéré :

$$V^{\pi}(s) = E\left(\sum_{i=1}^{\infty} \gamma^{t_i} r(s_i, t_i, \pi(s_i, t_i))\right)$$

Caractérisation des politiques optimales :

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

Equation d'optimalité

Hypothèses :

- r **est bornée,**
- r **est semi-continue supérieure,**
- **les durées de transition sont minorées par un réel strictement positif.**

Critère γ-pondéré :

$$V^{\pi}(s) = E\left(\sum_{i=1}^{\infty} \gamma^{t_i} r(s_i, t_i, \pi(s_i, t_i))\right)$$

Caractérisation des politiques optimales :

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

XMDP

Cadre de modélisation générique

SMDP+ et TMDP sont des XMDP

Résout le problème de l'action "attendre"

Equation d'optimalité

Quels algos d'optimisation des politiques ?

Algos spécifiques dédiés aux types de problèmes, ie. à l'écriture des fonctions p et r.

Plan

Problématique
- Exemples
- Représenter l'incertain et l'instationnaire

Modèles
- Modèles classiques
- SMDP, SMDP+ et TMDP
- XMDP, un modèle générique

Optimisation des politiques
- TMDPpoly - DP
- SMDP+ - BellErr

Conclusion

TMDPpoly - DP

Idée générale

On tire parti de :

- “attendre” est la seule action paramétrique
- “attendre” ne modifie pas l’état
- forme de p et r (L, P_μ, r)

Alternance d’actions de A et d’attentes (éventuellement nulles).

TMDPpoly - DP

Equation de Bellman

$$U(\mu,t) = \begin{cases} \int_{-\infty}^{\infty} P_\mu(t')[R(\mu,t,t') + V(s'_\mu,t')]dt' & \text{si } T_\mu = \text{ABS} \\ \int_{-\infty}^{\infty} P_\mu(t'-t)[R(\mu,t,t') + V(s'_\mu,t')]dt' & \text{si } T_\mu = \text{REL} \end{cases} \quad (1)$$

$$Q(s,t,a) = \sum_{\mu \in M} L(\mu|s,t,a) \cdot U(\mu,t) \quad (2)$$

$$\overline{V}(s,t) = \max_{a \in A} Q(s,t,a) \quad (3)$$

$$\tilde{V}(s,t) = \sup_{t' \geq t} \left(\int_t^{t'} K(s,\theta)d\theta + \overline{V}(s,t') \right) \quad (4)$$

$$V = S\tilde{V} \quad , S \text{ opérateur de réduction de degré} \quad (5)$$

TMDPpoly - DP

Difficultés

Passer à plus d'une variable continue

Comment écrire p ?

exemple : avancer pendant τ, $F(position' | position, avancer(\tau)) = ?$

$\longrightarrow$ Limitation de TMDPpoly dans l'expressivité du modèle

Plan

Problématique
- Exemples
- Représenter l'incertain et l'instationnaire

Modèles
- Modèles classiques
- SMDP, SMDP+ et TMDP
- XMDP, un modèle générique

Optimisation des politiques
- TMDPpoly - DP
- SMDP+ - BellErr

Conclusion

SMDP+ - BellErr

Idée générale

Partons d'une politique SMDP+ quelconque,
. . . on peut chercher la date à laquelle on peut le plus améliorer la politique courante
. . . et utiliser cette date pour améliorer le découpage de la droite temporelle.

On définit la *t-erreur de Bellman*.
On met en oeuvre un algorithme alternant optimisation d'une *politique discrète* et amélioration du découpage des *périodes de décision*.

L'idée : limiter la complexité par rapport à un problème stationnaire similaire en ne considérant que des périodes de décision utiles.

SMDP+ - BellErr

Définitions

Definition (Période de décision optimale)

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

Definition (t-erreur de Bellman)

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

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

SMDP+ - BellErr

Algorithme

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

SMDP+ - BellErr

Algorithme

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

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

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

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

SMDP+ - BellErr

Algorithme

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

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

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

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

Etape 3 : Recherche des dates de décision optimales.

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

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

SMDP+ - BellErr

Algorithme

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

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

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

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

Etape 3 : Recherche des dates de décision optimales.

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

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

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

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

SMDP+ - BellErr

Algorithme

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

$\longrightarrow$ Cache de dates de décision minimal

SMDP+ - BellErr

Algorithme

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

$\longrightarrow$ Cache de dates de décision minimal

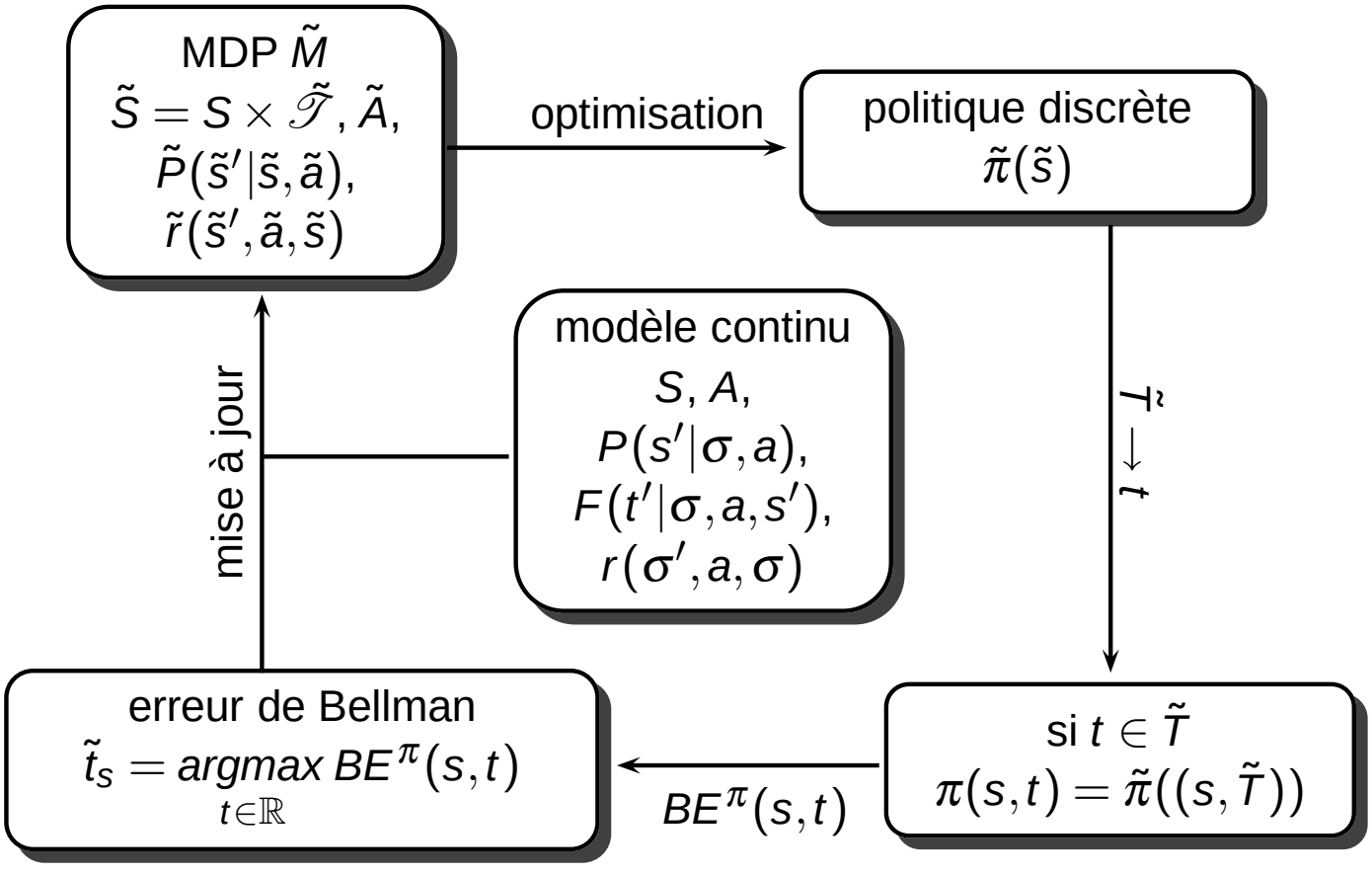
Condition d'arrêt :

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

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

SMDP+ - BellErr

Algorithme



SMDP+ - BellErr

Difficultés

Evaluation de la politique π avant calcul de l'erreur de Bellman

2 approches :

- $\longrightarrow$ approximation $\pi = \tilde{\pi}$
- $\longrightarrow$ approximation par projection sur une base de fonctions. Approche intéressante car :
  - similaire à la projection sur les polynômes de TMDPpoly
  - similaire à la projection sur une base finie de fonctions des méthodes ALP
  - challenge : comment exploiter l'info $\pi \approx \tilde{\pi}$ lors de la projection ?

Conclusion

Travaux en cours

- TMDPpoly $\longrightarrow$ debuggage
- Ecriture de problèmes réels pour adaptation de TMDP / SMDP+
- Ouverture vers MDP à espaces d'états factorisés hybrides et ALP.

Merci de votre attention !