

Un Algorithme Amélioré d'Itération de la Politique Approchée pour les Processus Décisionnels Semi-Markoviens Généralisés

Emmanuel Rachelson ¹

Patrick Fabiani ¹
Frédéric Garcia ²

¹ONERA-DCSD

²INRA-BIA

JFPDA08, 20 juin 2008

Plan

Problèmes de Markov temporels

Exemples et motivation

Formalisation - le lien avec les MDP

Complexité du processus temporel

Méthode de résolution

Apprentissage de π à partir de simulations

L'algorithme online-ATPI

Résultats sur le problème du métro

Oui mais ...

Fragilité de l'algorithme

online-ATPI amélioré

Travaux actuels

Plan

Problèmes de Markov temporels

Exemples et motivation

Formalisation - le lien avec les MDP

Complexité du processus temporel

Méthode de résolution

Apprentissage de π à partir de simulations

L'algorithme online-ATPI

Résultats sur le problème du métro

Oui mais ...

Fragilité de l'algorithme

online-ATPI amélioré

Travaux actuels



Exemples et motivation

Décision dans l'incertain avec forte dynamique temporelle.

→ planifier pour se coordonner avec un environnement instationnaire et incertain.

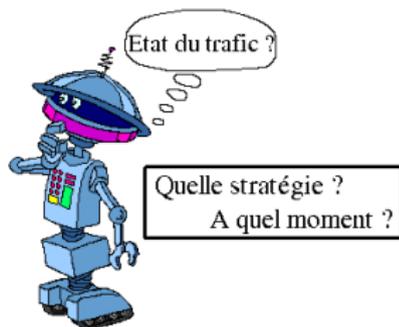
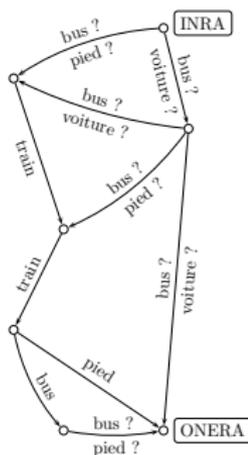


Exemples et motivation

Décision dans l'incertain avec forte dynamique temporelle.

→ planifier pour se coordonner avec un environnement instationnaire et incertain.

Aller de l'INRA à l'ONERA ?





Exemples et motivation

Décision dans l'incertain avec forte dynamique temporelle.

→ planifier pour se coordonner avec un environnement instationnaire et incertain.

Quand ouvrir quel guichet ?



Exemples et motivation

Décision dans l'incertain avec forte dynamique temporelle.

→ planifier pour se coordonner avec un environnement instationnaire et incertain.

Routage des avions sur taxiway





Exemples et motivation

Décision dans l'incertain avec forte dynamique temporelle.

→ planifier pour se coordonner avec un environnement instationnaire et incertain.

Coordination à bord



Exemples et motivation

Décision dans l'incertain avec forte dynamique temporelle.

→ planifier pour se coordonner avec un environnement instationnaire et incertain.

Ajouter ou retirer des rames ?



Caractéristiques des exemples

Pourquoi écrire un MDP pour les exemples précédents est-il une tâche difficile ?

→ “il se passe beaucoup de choses complexes en parallèle”

- dynamique partiellement incontrôlable
- phénomènes concurrents



Concurrence et (S)MDP

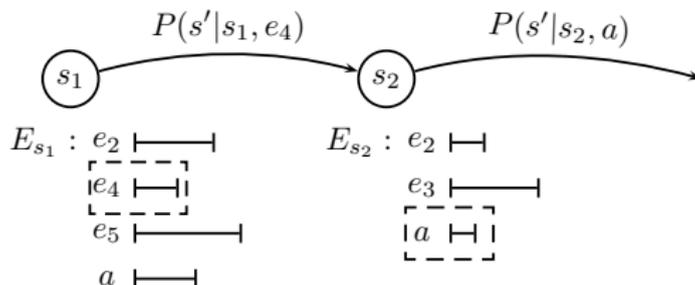
GSMDP, (Younes et al., 04)

Plusieurs processus semi-Markoviens affectant le même état

→ GSMP, (Glynn, 89)

+ choix d'actions.

→ $\langle S, E, A, P, F, r \rangle$



Propriétés des GSMDP

Se ramènent à des MDP si les horloges sont observables ...
pb : les horloges ne sont pas observables.
→ dynamique non-Markovienne.

(📄 *Younes et al., 04*) : approximation basée sur les distributions
phase-type.
Introduit des états supplémentaires.

Proposition : Apprendre $\pi(s, t)$ à partir d'un modèle générateur.

Contrôler un GSMDP

→ définir une politique, oui, mais sur quel espace d'observations ?

- *état naturel* - processus non-Markovien, pas de garantie d'optimalité
- *état naturel* + horloges - Markovien mais ... inobservable en pratique (📄 Nilsen, 98)
- *état naturel* + durées d'activation + états d'activation - Markovien, observable ... mais fait exploser la dimension de l'espace des observations ($|S|(|E| + 1) + |E|$ variables).

Notre choix : $\pi(s)$.

Caractéristiques des problèmes traités

- Temps observable borné, $\gamma = 1$
- Longues trajectoires entre $t = 0$ et $t = t_f$
- Espaces d'états hybrides de grande dimension

Plan

Problèmes de Markov temporels

Exemples et motivation

Formalisation - le lien avec les MDP

Complexité du processus temporel

Méthode de résolution

Apprentissage de π à partir de simulations

L'algorithme online-ATPI

Résultats sur le problème du métro

Oui mais ...

Fragilité de l'algorithme

online-ATPI amélioré

Travaux actuels

L'idée générale

On suppose qu'on dispose d'un modèle générateur / d'un simulateur

Jouer la politique

⇔ tirer des trajectoires

⇔ ensemble fini de réalisations de la var. aléa. $V^\pi(s)$

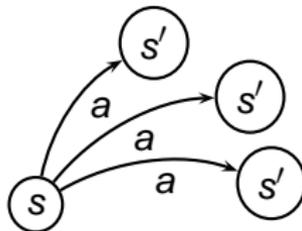
⇒ fournit une estimation de $V^\pi(s)$

travaux similaires : roll-out ( *Kearns et al., 02*), Monte-Carlo, ...

Idée : amélioration à un coup, itérative et locale de π .

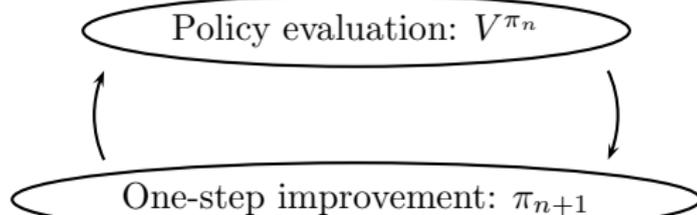
En d'autres termes ...

$$[Q(s, a)]_k = E(r(s, a, s')) + E(\tilde{V}(s'))$$



online-ATPI

- Itération de la politique approchée
- Phase d'évaluation ? Simulation de la politique
- Généralisation de la fonction de valeur par apprentissage à partir des simulations
- Stockage de la politique ? Instanciation en ligne



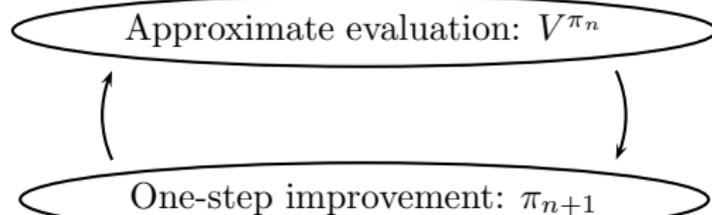
online-ATPI

- Itération de la politique approchée
- Phase d'évaluation ? Simulation de la politique
- Généralisation de la fonction de valeur par apprentissage à partir des simulations
- Stockage de la politique ? Instanciation en ligne

Intérêt : recherche monotone dans l'espace des politiques,
on évite les minima locaux.
→ bon comportement anytime.

online-ATPI

- Itération de la politique approchée
- Phase d'évaluation ? Simulation de la politique
- Généralisation de la fonction de valeur par apprentissage à partir des simulations
- Stockage de la politique ? Instanciation en ligne



online-ATPI

- Itération de la politique approchée
- Phase d'évaluation ? Simulation de la politique
- Généralisation de la fonction de valeur par apprentissage à partir des simulations
- Stockage de la politique ? Instanciation en ligne

Plus précisément : asynchronous API ... et online-API.

online-ATPI

- Itération de la politique approchée
- Phase d'évaluation ? Simulation de la politique
- Généralisation de la fonction de valeur par apprentissage à partir des simulations
- Stockage de la politique ? Instanciation en ligne

Echantillonnage dans l'espace des trajectoires

online-ATPI

- Itération de la politique approchée
- Phase d'évaluation ? Simulation de la politique
- Généralisation de la fonction de valeur par apprentissage à partir des simulations
- Stockage de la politique ? Instanciation en ligne

→ Méthode utilisée : méthodes à noyaux.

online-ATPI

- Itération de la politique approchée
- Phase d'évaluation ? Simulation de la politique
- Généralisation de la fonction de valeur par apprentissage à partir des simulations
- Stockage de la politique ? Instanciation en ligne

→ Calcul de l'action optimale par roll-out et échantillonnage sur un coup en utilisant la fonction de valeur apprise

online-ATPI

main :

Input : π_0 or \tilde{V}_0, s_0

loop

$TrainingSet \leftarrow \emptyset$

for $i = 1$ to N_{sim} **do**

$\{(s, v)\} \leftarrow \text{simulate}(\tilde{V}, s_0)$

$TrainingSet \leftarrow TrainingSet \cup \{(s, v)\}$

end for

$\tilde{V} \leftarrow \text{TrainApproximator}(TrainingSet)$

end loop



online-ATPI

simulate(\tilde{V}, s_0) :

$ExecutionPath \leftarrow \emptyset$

$s \leftarrow s_0$

while horizon not reached **do**

$action \leftarrow \text{ComputePolicy}(s, \tilde{V})$

$(s', r) \leftarrow \text{GSMDPstep}(s, action)$

$ExecutionPath \leftarrow ExecutionPath \cup (s', r)$

end while

convert execution path to value function $\{(s, v)\}$

return $\{(s, v)\}$

online-ATPI

ComputePolicy(s, \tilde{V}) :

for $a \in A$ **do**

$$\tilde{Q}(s, a) = 0$$

for $j = 1$ to N_{samples} **do**

$$(s', r) \leftarrow \text{GSMDPstep}(s, a)$$

$$\tilde{Q}(s, a) \leftarrow \tilde{Q}(s, a) + r + \gamma^{t'-t} \tilde{V}(s')$$

end for

$$\tilde{Q}(s, a) \leftarrow \frac{1}{N_{\text{samples}}} \tilde{Q}(s, a)$$

end for

$$\text{action} \leftarrow \arg \max_{a \in A} \tilde{Q}(s, a)$$

return *action*

Résultats d'optimisation

	π_0	π_1	π_2	π_3	π_4
t_{sim}	47.1	203.43	206.45	446.15	1504.41
t_{learn}	2.28	2.7	12.18	56.08	229.45
$\tilde{V}_{stat}(s_0)$	-3261.31	-3188.11	-2074.74	-1850.12	-887.076
$\tilde{V}_{SVM}(s_0)$	-2980.29	-2962.46	-2020.22	-1837.41	-875.417
#SV	55	61	439	3588	13596

Plan

Problèmes de Markov temporels

Exemples et motivation

Formalisation - le lien avec les MDP

Complexité du processus temporel

Méthode de résolution

Apprentissage de π à partir de simulations

L'algorithme online-ATPI

Résultats sur le problème du métro

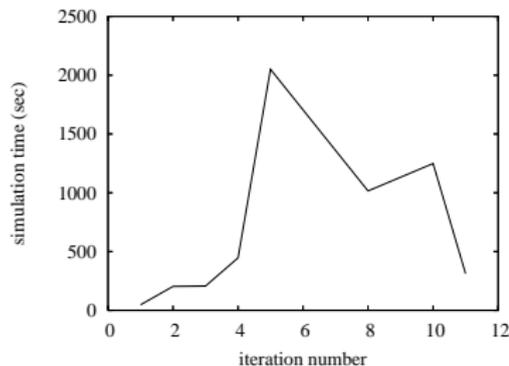
Oui mais ...

Fragilité de l'algorithme

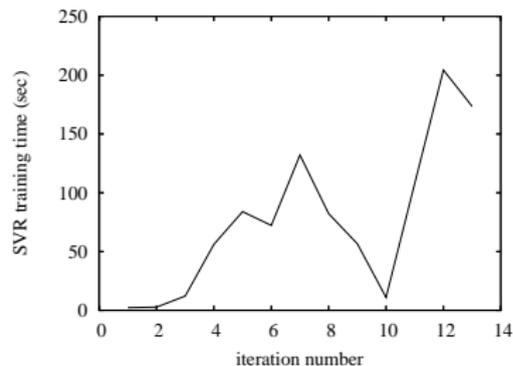
online-ATPI amélioré

Travaux actuels

Résultats d'optimisation cont'd



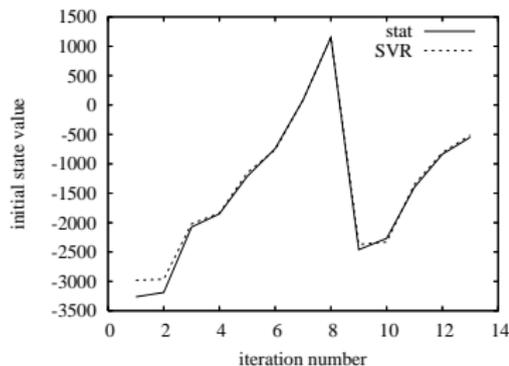
(a) Simulation Time



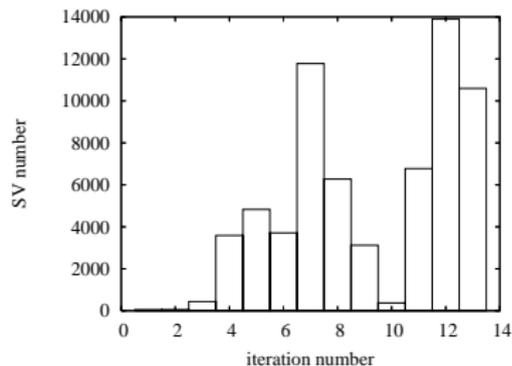
(b) SVR Training Time



Résultats d'optimisation cont'd



(c) Policy quality



(d) Number of support vectors

Causes

Un problème de convergence et de monotonie dû à :

- l'erreur d'approximation dans API
- le traitement "memoryless" des zones déjà explorées de S



Illustration graphique

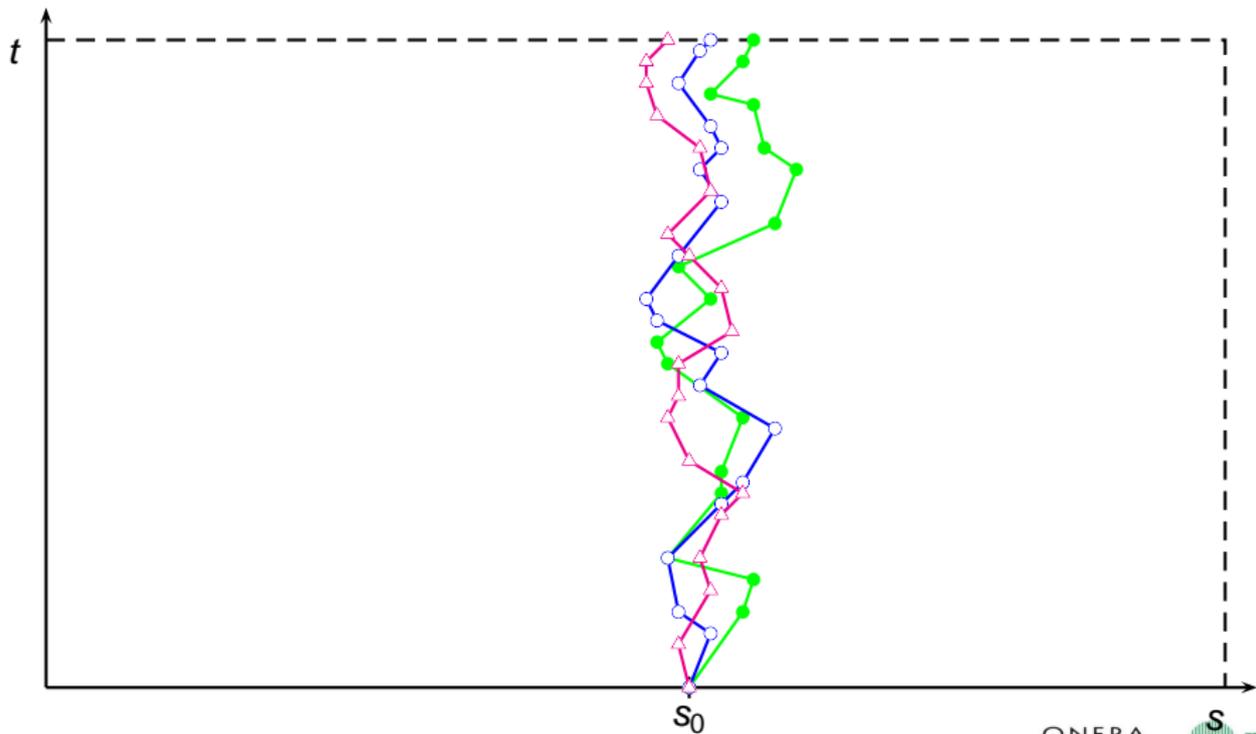


Illustration graphique

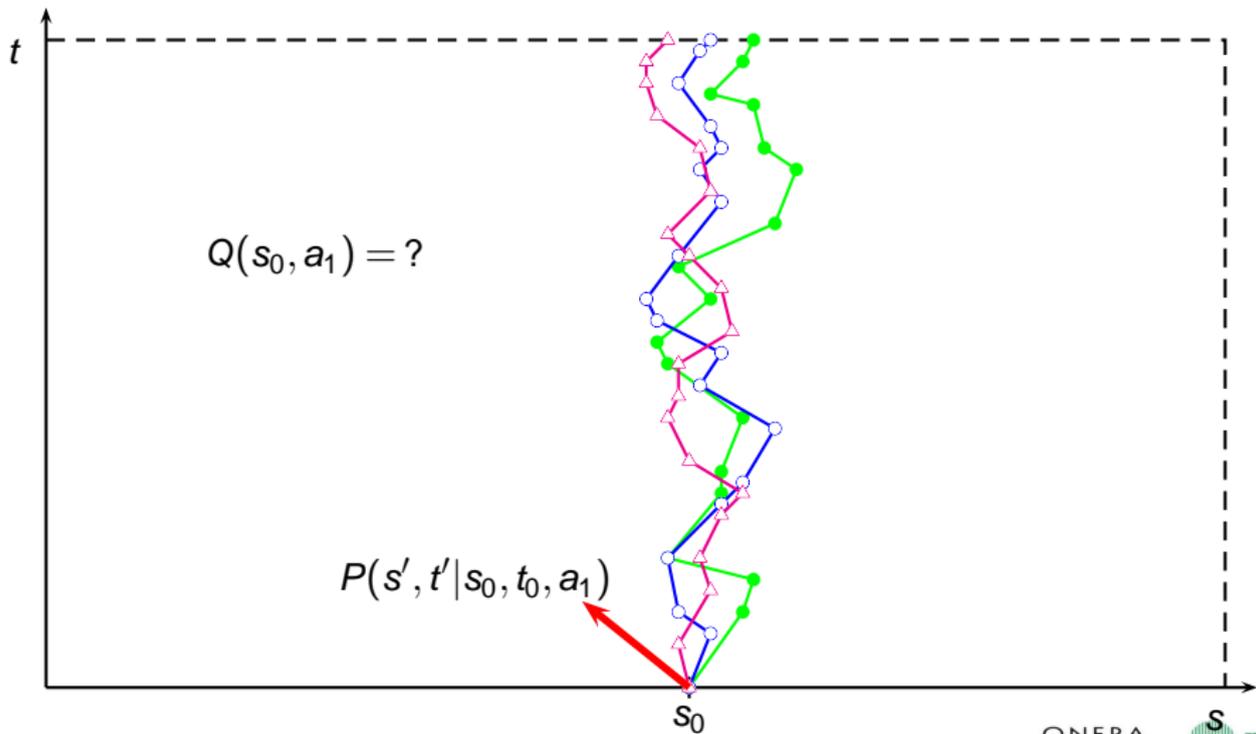




Illustration graphique

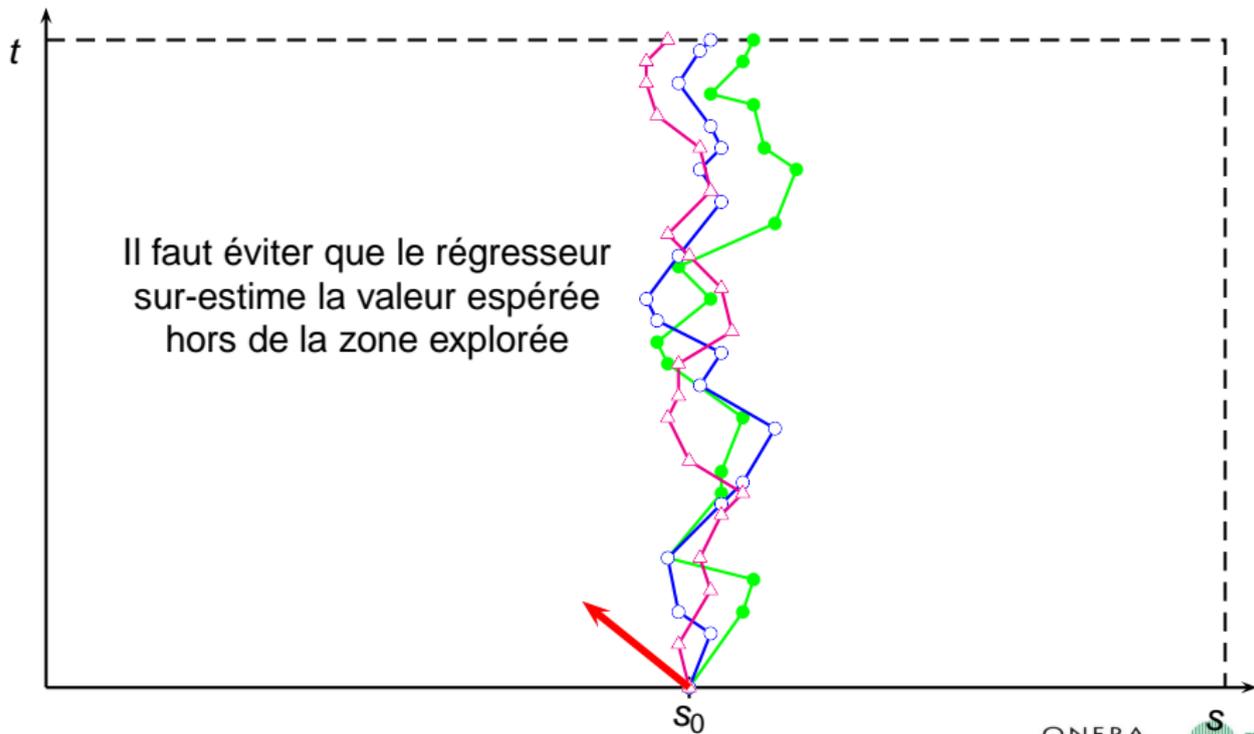
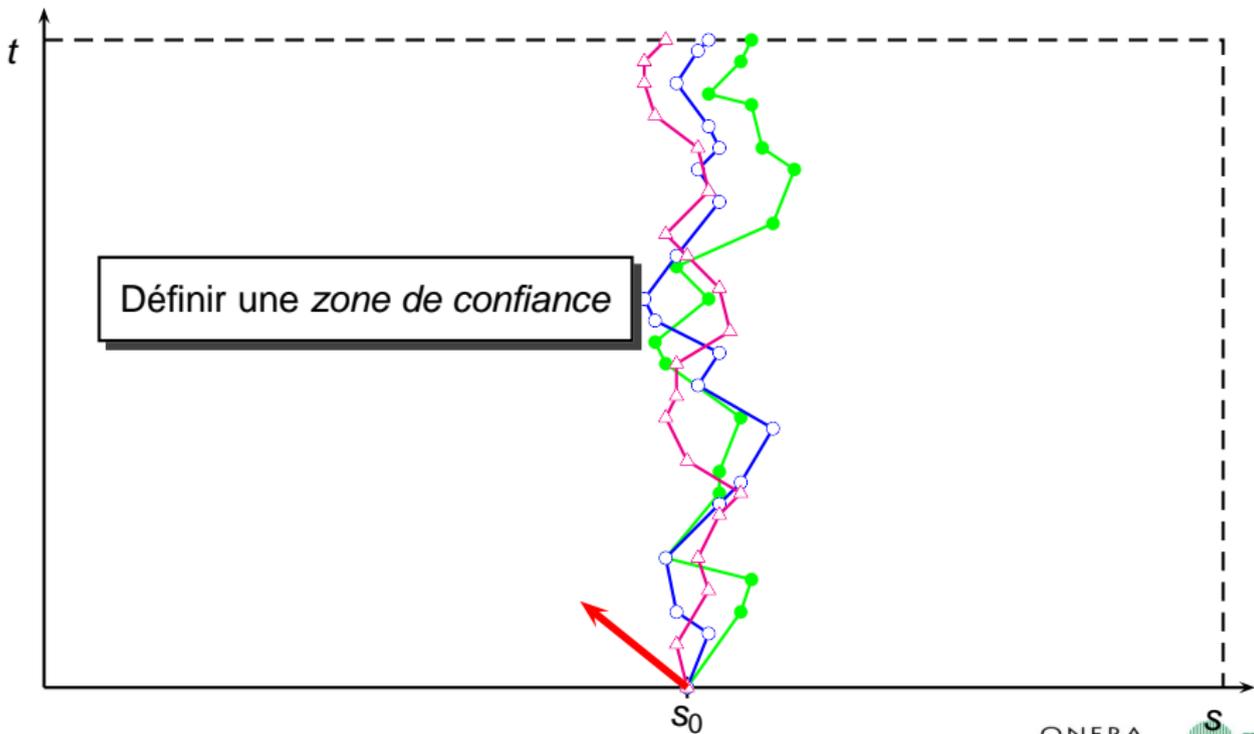




Illustration graphique



online-ATPI amélioré

- A quoi correspond la zone de confiance
→ s'approcher de la statistique suffisante pour $V^\pi(s)$
On l'approche par une estimation de densité.
- Utiliser la zone de confiance
→ que fait-on hors de cette zone ?
simuler ? quelle politique ?
- Stocker une politique ?

Pb de régression, de classification et d'estimation de densité.
→ la fonction *simulateWithStop* remplace *simulate*.

Mise en perspective

Utilisation des techniques d'apprentissage statistique pour la recherche incrémentale de politiques.

→ à mettre en parallèle de ( *Lagoudakis et al., 03*)

Optimalité ?

→ handicapée par l'exploration partielle de l'espace d'états.

Compromis exploration/exploitation : définition d'une politique d'exploration hors zone de confiance.

Comment tout résumer en un tableau

Objectifs actuels :

- comparer les différents modes de calcul et représentation de
 - π : base de données, SVC
 - V : base de données, SVR, kLASSO, LWPR.
 - $conf$: base de données, OC-SVM, LWPR.
- comparer plusieurs politiques d'exploration
 - compromis exploitation / aléatoire / heuristique basée sur $conf$
- comparer les approches avec stockage de π seul ($pb0$), celles avec V seule ($pb1$) et celles avec V et π ($pb2$).

Comment tout résumer en un tableau

	$\tilde{\pi}$	$\tilde{\pi} / \pi_{alea} / \pi_{heurist}$
$pb0 (\tilde{\pi})$		
$pb1 (\tilde{V})$		
$pb2 (\tilde{\pi} \text{ et } \tilde{V}^\pi)$		

Bibliothèque C++ GiSMoP

→ <http://emmanuel.rachelson.free.fr/fr/gismop.html>

Conclusion

GSMDP Une méthode pour le “passage à l'échelle” des problèmes stochastiques temporels ; combine des résultats de

- modélisation en processus décisionnels stochastiques,
- apprentissage par renforcement,
- apprentissage statistique

API Une généralisation de l'itération de la Politique Approchée à

- l'exploration partielle et incrémentale
- la régression statistique

Résultats plus complets de GiSMoP à EWRL08.