Une Approche basée sur la Simulation pour l'Optimisation des Processus Décisionnels Semi-Markoviens Généralisés

Emmanuel Rachelson [1]

Patrick Fabiani [1]
Frédérick Garcia [2]
Gauthier Quesnel [2]

[1]ONERA-DCSD

[2]INRA-BIA

CAp08, 30 mai 2008

ONERA THE FRENCH AEROSPACE LAB INRA

Plan

Plan

Exemples et motivation

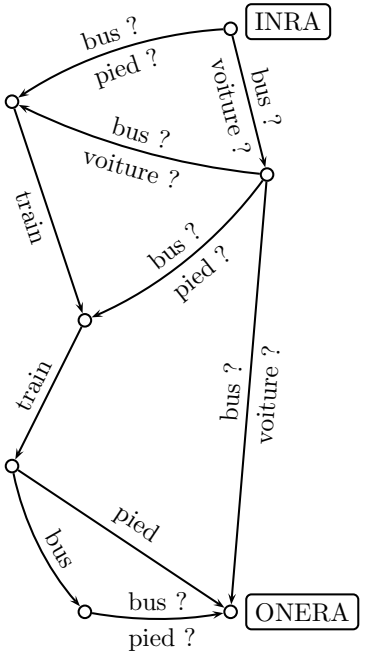
Décision dans l'incertain avec forte dynamique temporelle.

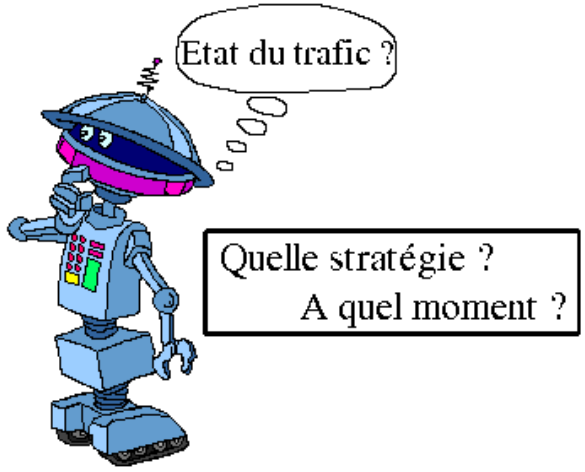
$\longrightarrow$ planifier pour se coordonner avec un environnement instationnaire et incertain.

Exemples et motivation

Décision dans l'incertain avec forte dynamique temporelle.
$\longrightarrow$ 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.
$\longrightarrow$ 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.
$\longrightarrow$ planifier pour se coordonner avec un environnement instationnaire et incertain.

Routage des avions sur taxiway

ONERA THE FRENCH AEROSPACE LAB INRA

Exemples et motivation

Décision dans l'incertain avec forte dynamique temporelle.
$\longrightarrow$ planifier pour se coordonner avec un environnement instationnaire et incertain.

Coordination à bord

Exemples et motivation

Décision dans l'incertain avec forte dynamique temporelle.
$\longrightarrow$ 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 ?

$\longrightarrow$ "il se passe beaucoup de choses complexes en parallèle"

- dynamique partiellement incontrolable
- phénomènes concurrents

Concurrence et (S)MDP

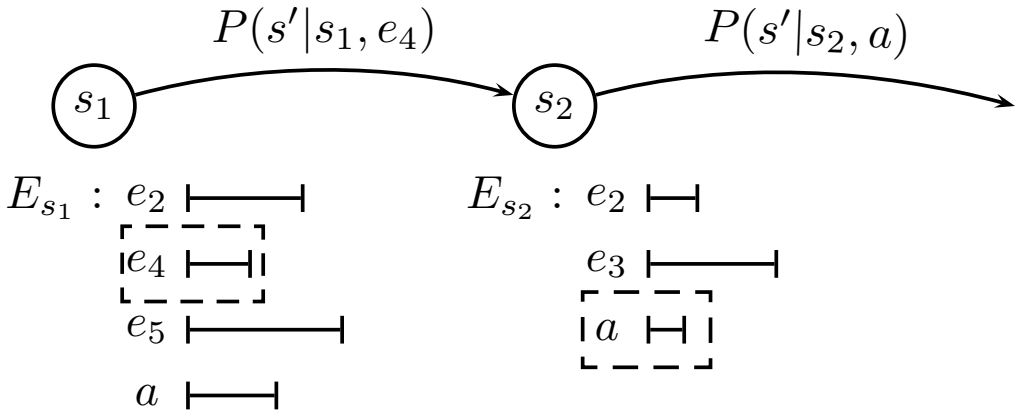
GSMDP, (*Younes et al., 04)*

Plusieurs processus semi-Markoviens affectant le même état

$\longrightarrow$ GSMP, (*Glynn, 89)*

+ choix d'actions.

$\longrightarrow \langle S, E, A, P, F, r \rangle$



Propriétés des GSMDP

Se ramènent à des MDP si les horloges sont observables ...
pb : les horloges ne sont pas observables.
$\longrightarrow$ 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

$\longrightarrow$ 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

L'idée générale

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

Jouer la politique

$\Leftrightarrow$ tirer des trajectoires

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

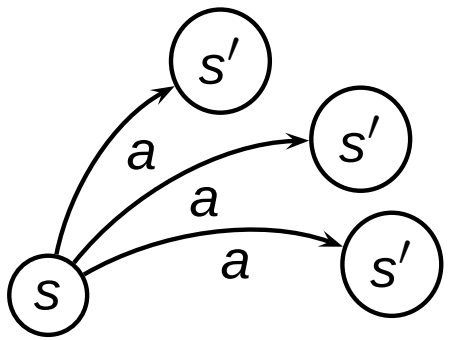
$\Rightarrow$ fournit une estimation de $V^{\pi}(s)$

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

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

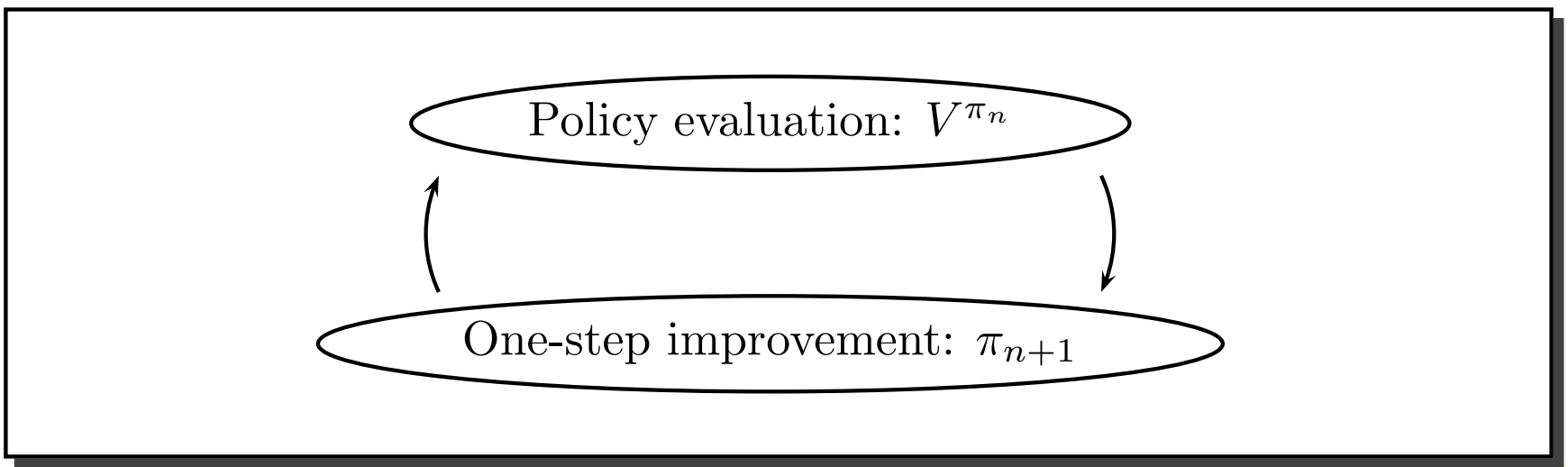
En d'autres termes ...

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



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



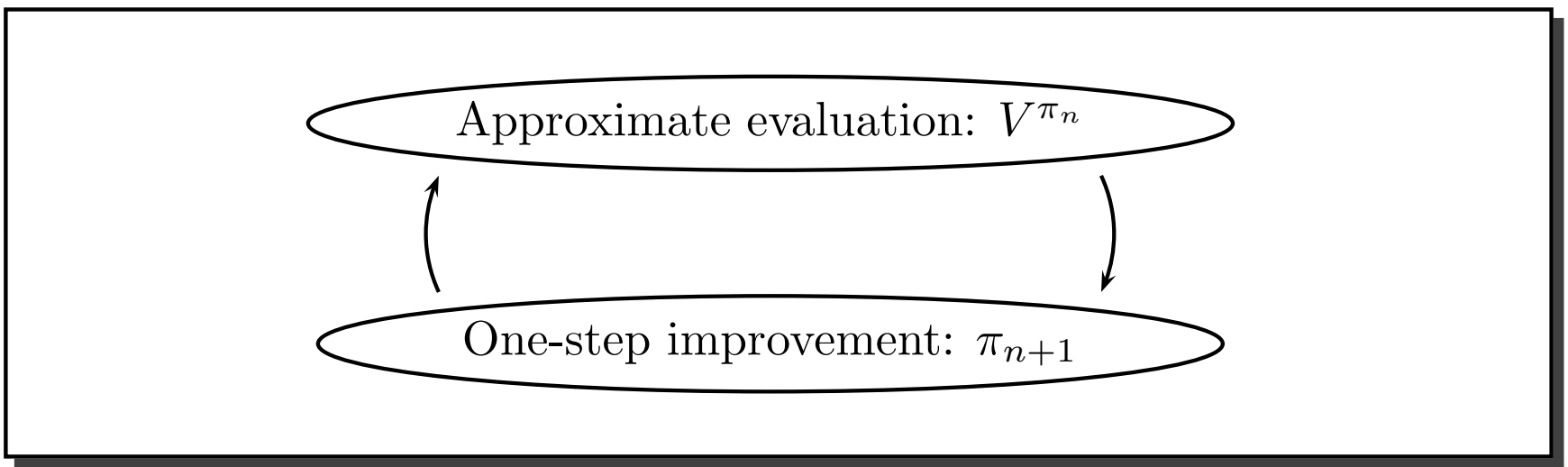
online-ATPI

- 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.
$\longrightarrow$ 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

Echantillonage 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

$\longrightarrow$ Méthode utilisée : SVR.

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

$\longrightarrow$ Calcul de l'action optimale par roll-out et échantillonage 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$(s, \tilde{V})$ **:**

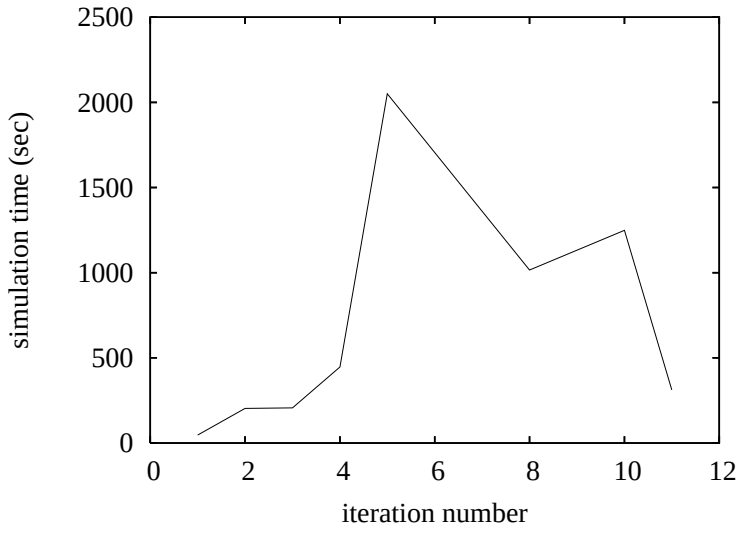
for $a \in A$ **do**
  $\tilde{Q}(s,a) = 0$
  for $j = 1$ to $N_{samples}$ **do**
    $(s', r) \leftarrow \text{GSMDPstep}(s,a)$
    $\tilde{Q}(s,a) \leftarrow \tilde{Q}(s,a) + r + \gamma^{t'-t}\tilde{V}(s')$
  end for
  $\tilde{Q}(s,a) \leftarrow \frac{1}{N_{samples}}\tilde{Q}(s,a)$
end for
$action \leftarrow \arg\max_{a \in A} \tilde{Q}(s,a)$
return *action*

Résultats d'optimisation

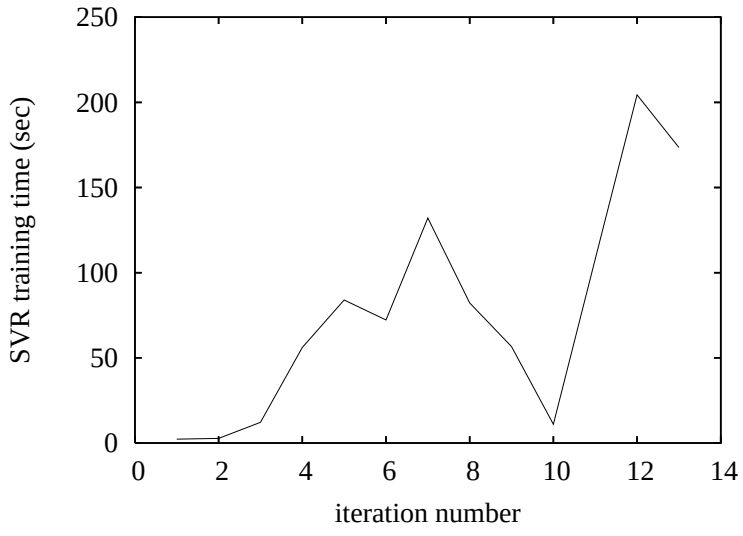
	π_0	π_1	π_2	π_3	π_4
t_{sim}	47.1	203.43	206.45	446.15	1504.41
t_{learn}	2.28	2.7	12.18	56.08	229.45
$\tilde{V}_{stat}(s_0)$	-3261.31	-3188.11	-2074.74	-1850.12	-887.076
$\tilde{V}_{SVM}(s_0)$	-2980.29	-2962.46	-2020.22	-1837.41	-875.417
#SV	55	61	439	3588	13596

Plan

Résultats d'optimisation cont'd

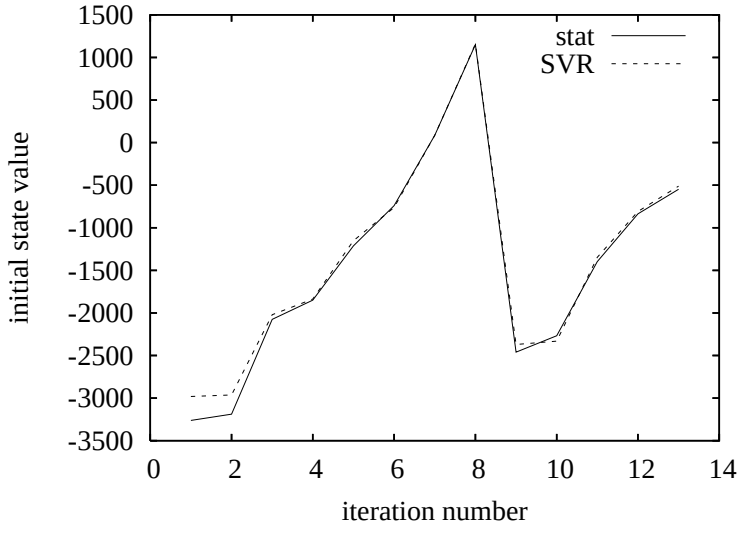


(a) Simulation Time

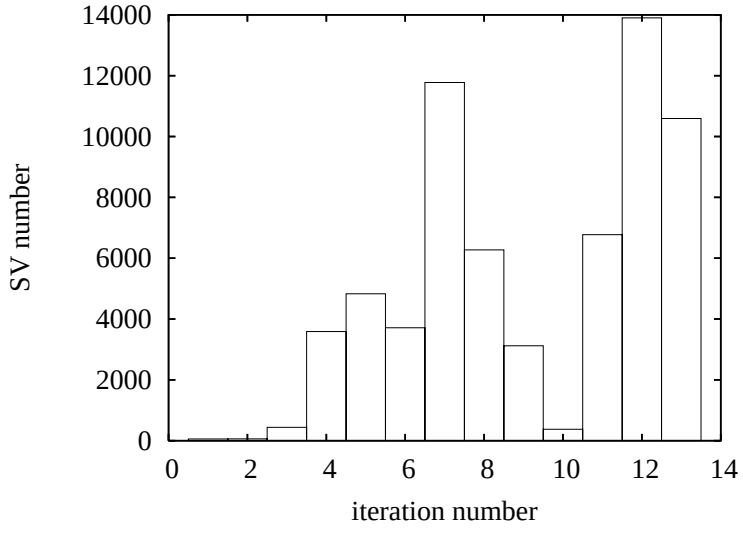


(b) SVR Training Time

Résultats d'optimisation cont'd



(c) Policy quality



(d) Number of suport vectors

Illustration graphique

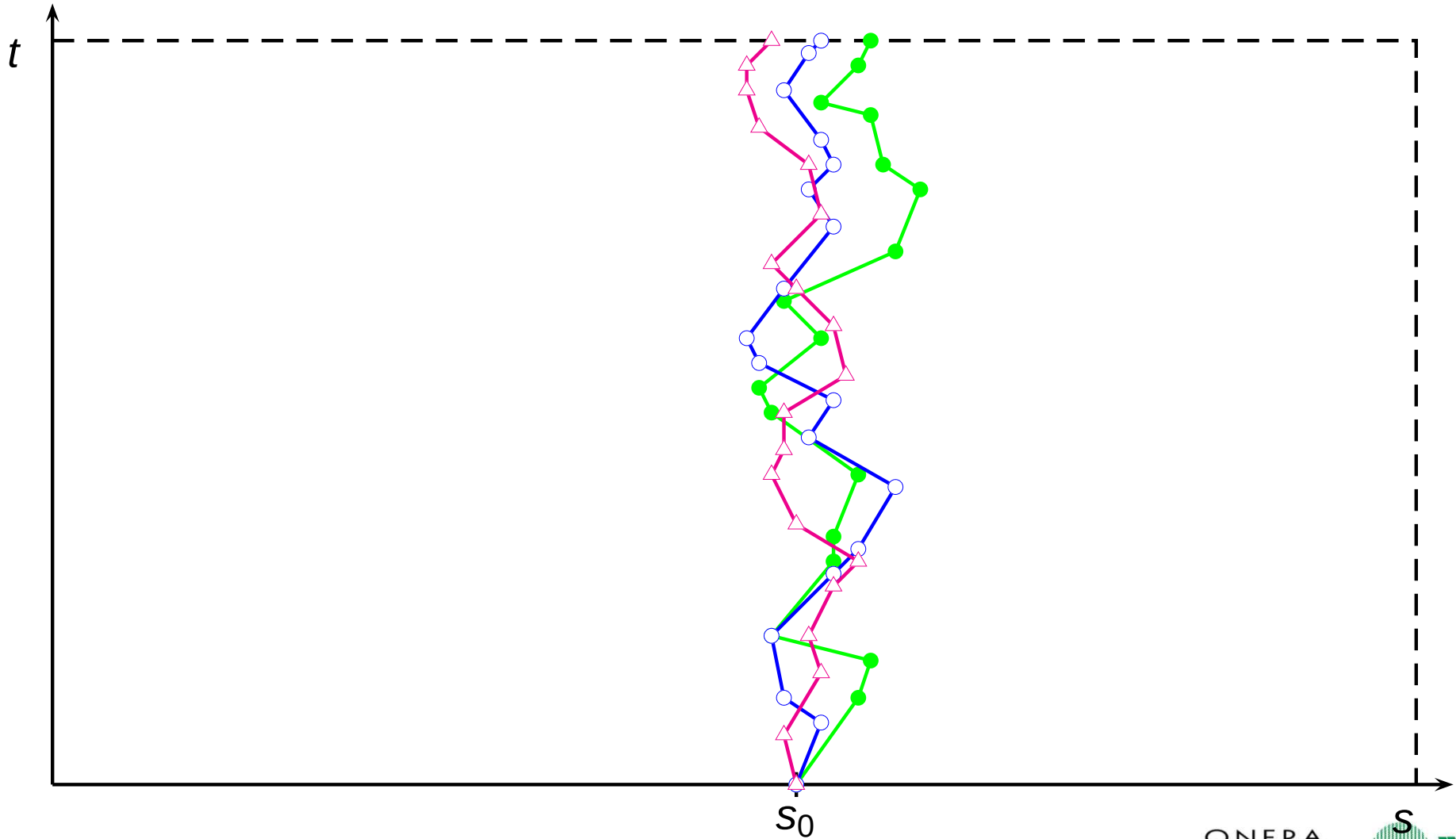


Illustration graphique

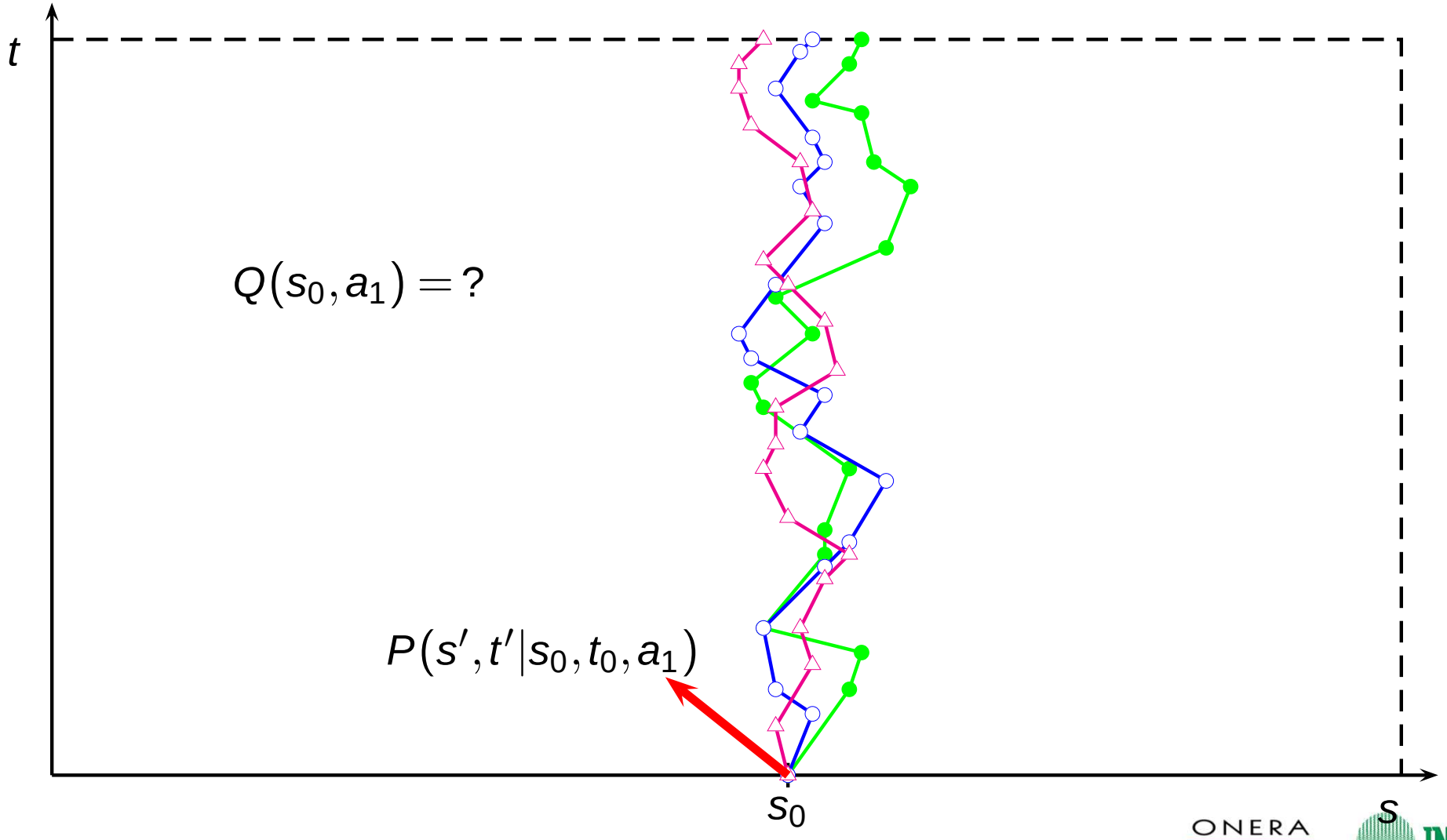
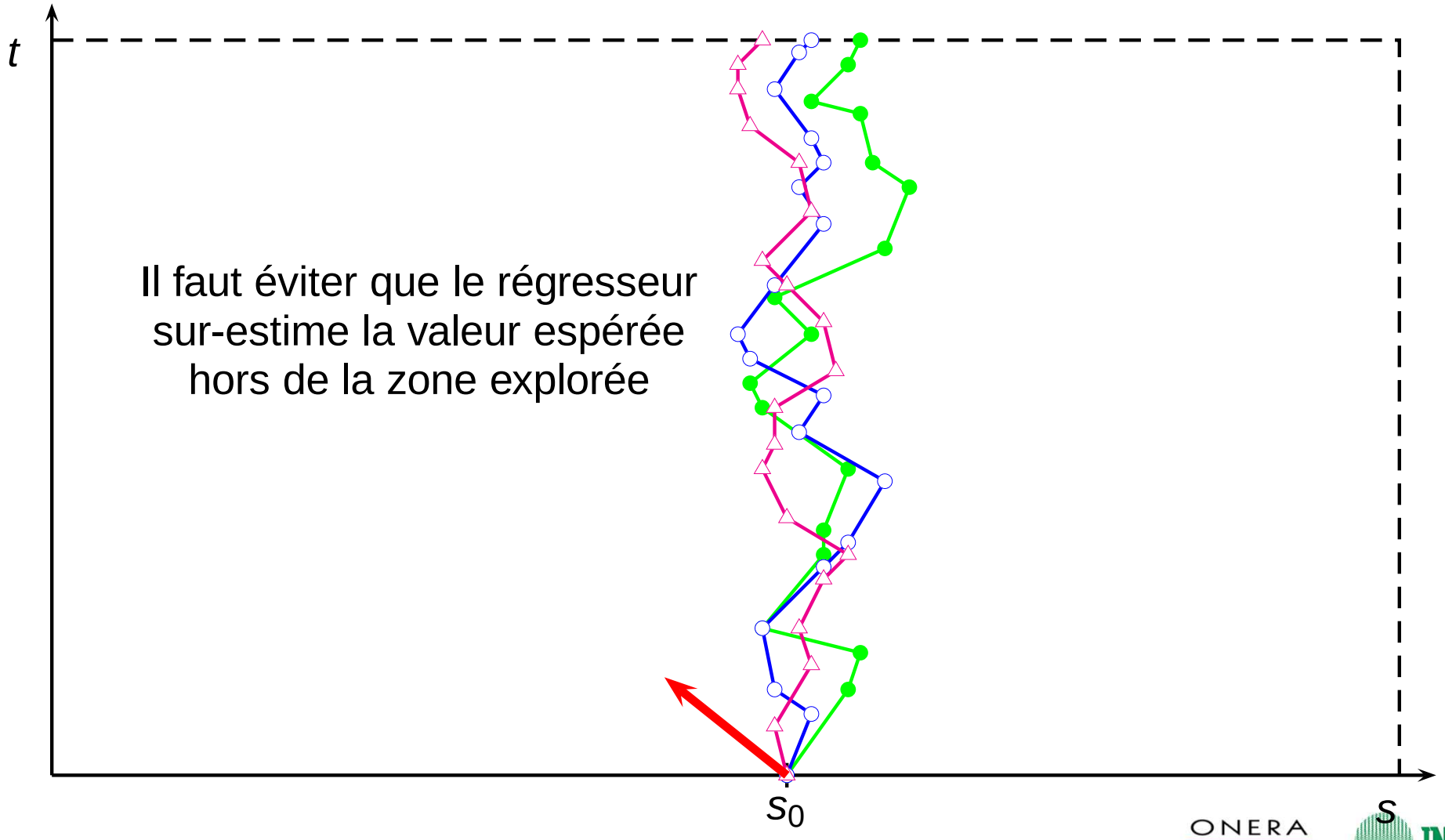


Illustration graphique



Conclusion

- 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
- Des résultats bons à condition d'utiliser une politique initiale suffisamment informative
- Un problème de convergence et de monotonie dû à :
  - l'erreur d'approximation dans API
  - le traitement "memoryless" des zones déjà explorées de S
- . . . problèmes abordés dans une version améliorée de l'algorithme (EWRL08)