Simulation-based Approximate Policy Iteration for Generalized Semi-Markov Decision Processes

Emmanuel Rachelson [1]

Patrick Fabiani [1]
Frédérick Garcia [2]

[1]ONERA-DCSD

[2]INRA-BIA
Toulouse, France

EWRL08, July 3rd, 2008

Plan

1. Temporal Markov Problems: motivation and modeling
   - Examples
   - Problem features
   - GSMDP
2. Solving large scale GSMDP: ATPI
   - Basic ideas
   - Introducing confidence
   - The bigger picture

Plan

1. Temporal Markov Problems: motivation and modeling
   - Examples
   - Problem features
   - GSMDP
2. Solving large scale GSMDP: ATPI
   - Basic ideas
   - Introducing confidence
   - The bigger picture

Planning under uncertainty with time dependency.
$\longrightarrow$ planning to coordinate with an uncertain and unstationnary environment.

Planning under uncertainty with time dependency.
$\longrightarrow$ planning to coordinate with an uncertain and unstationnary environment.

Should we open more lines ?

Planning under uncertainty with time dependency.
$\longrightarrow$ planning to coordinate with an uncertain and unstationnary environment.

Airplanes taxiing management

Planning under uncertainty with time dependency.
$\longrightarrow$ planning to coordinate with an uncertain and unstationnary environment.

Onboard planning for coordination

Planning under uncertainty with time dependency.
$\longrightarrow$ planning to coordinate with an uncertain and unstationnary environment.

Adding or removing trains ?

Subway problem: toy example

Some figures

4 trains, 6 stations
$\longrightarrow$ 22 state variables, 9 actions

episodes of 12 hours with around 2000 steps.

Main idea

Why is writing an MDP for the previous problems such a difficult task ?

"Lots of things occur in parallel"

- concurrent phenomena
- partially controlable dynamics

Main idea

Why is writing an MDP for the previous problems such a difficult task ?

"Lots of things occur in parallel"

- concurrent phenomena
- partially controlable dynamics

Main idea

Why is writing an MDP for the previous problems such a difficult task ?

"Lots of things occur in parallel"

- concurrent phenomena
- partially controlable dynamics

Typical features

- Continuous time
- Hybrid state spaces
- Large state spaces
- Total reward criteria
- Long trajectories

How do we model all this ?

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

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

GSMP, (*Glynn, 89*)
Several semi-Markov processes affecting the same state space

One process conditionned by the choice of the action undertaken

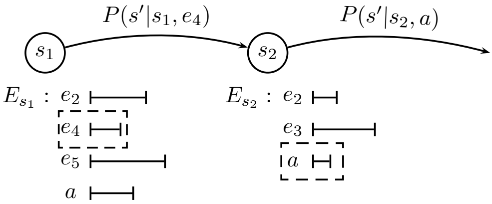
$$\rightarrow \langle S, E, A, P, F, r \rangle$$

GSMDP, (Younes et al., 04)

GSMP, (Glynn, 89)
Several semi-Markov processes affecting the same state space

One process conditionned by the choice of the action undertaken

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



Controling GSMDP

non-Markov behaviour !

$\longrightarrow$ no guarantee of an optimal Markov policy

(*Younes et al., 04*): approximate your model with phase-type (exponential) distributions.

Supplementary variables technique (*Nilsen, 98*).

Our approach: no hypothesis, simulation-based API.

Controling GSMDP

non-Markov behaviour !

$\longrightarrow$ no guarantee of an optimal Markov policy

(*Younes et al., 04*): approximate your model with phase-type (exponential) distributions.

Supplementary variables technique (*Nilsen, 98*).

Our approach: no hypothesis, simulation-based API.

Controling GSMDP

non-Markov behaviour !

$\longrightarrow$ no guarantee of an optimal Markov policy

(*Younes et al., 04*): approximate your model with phase-type (exponential) distributions.

Supplementary variables technique (*Nilsen, 98*).

Our approach: no hypothesis, simulation-based API.

Plan

1. Temporal Markov Problems: motivation and modeling
   - Examples
   - Problem features
   - GSMDP

2. Solving large scale GSMDP: ATPI
   - Basic ideas
   - Introducing confidence
   - The bigger picture

Contribution overview

General framework:

- API, simulation-based PI.

Our contribution:

- API as non-parametric statistical learning:
  - classification (policy),
  - regression (value function),
  - density estimation (“I don’t know” situations)
- Three extensive uses of simulation:
  - Monte-Carlo sampling for the evaluation of V^{π}
  - Roll-out for the calculation of Q-values
  - Selection of the subset of states on which we perform policy improvement

Simulation-based policy evaluation

Our hypothesis: we have a generative model of the process.
$\longrightarrow$ (Monte-Carlo) simulation-based policy evaluation.

Statistical learning

Simulating the policy
$\Leftrightarrow$ Drawing a set of *trajectories*
$\Leftrightarrow$ Finite set of realisations of r.v. $R^{\pi}(s)$

We need to

- abstract (*generalize*) information from samples
- *compactly* store previous knowledge of $V^{\pi}(s) = E(R^{\pi}(s))$.

(nearest neighbours, SVR, kLASSO, LWPR)

Simulation-based policy evaluation

Our hypothesis: we have a generative model of the process.

$\longrightarrow$ (Monte-Carlo) simulation-based policy evaluation.

Statistical learning

Simulating the policy

$\Leftrightarrow$ Drawing a set of *trajectories*

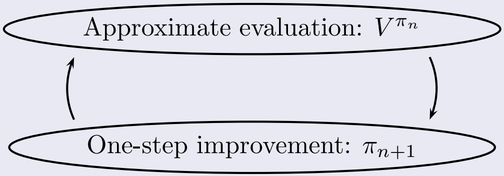
$\Leftrightarrow$ Finite set of realisations of r.v. $R^\pi(s)$

We need to

- abstract (*generalize*) information from samples
- *compactly* store previous knowledge of $V^\pi(s) = E(R^\pi(s))$.

(nearest neighbours, SVR, kLASSO, LWPR)

Approximate Policy Iteration



Approximate Policy Iteration

in each visited state: 1-step rollout in order to find the best Q-value.
$\longrightarrow$ local improvements guided by the simulation of π_{n+1}.

online-API, cont'd

Motivation: don't want / can't improve the policy everywhere

- too time/resource consuming
- not useful with regard to 'relevant' information gathered

Useful ? Interesting ? Relevant ?
→ "Improving the policy in the situations I am likely to encounter today"

In other words . . .

Which subset of states for API ?

The ones visited by policy simulation !

online-API, cont'd

Motivation: don't want / can't improve the policy everywhere

- too time/resource consuming
- not useful with regard to 'relevant' information gathered

Useful ? Interesting ? Relevant ?
$\longrightarrow$ "Improving the policy in the situations I am likely to encounter today"

In other words . . .

Which subset of states for API ?

The ones visited by policy simulation !

online-API, cont'd

Motivation: don't want / can't improve the policy everywhere

- too time/resource consuming
- not useful with regard to 'relevant' information gathered

Useful ? Interesting ? Relevant ?
$\longrightarrow$ "Improving the policy in the situations I am likely to encounter today"

In other words . . .

Which subset of states for API ?

The ones visited by policy simulation !

online-API, cont'd

Motivation: don't want / can't improve the policy everywhere

- too time/resource consuming
- not useful with regard to 'relevant' information gathered

Useful ? Interesting ? Relevant ?
$\longrightarrow$ "Improving the policy in the situations I am likely to encounter today"

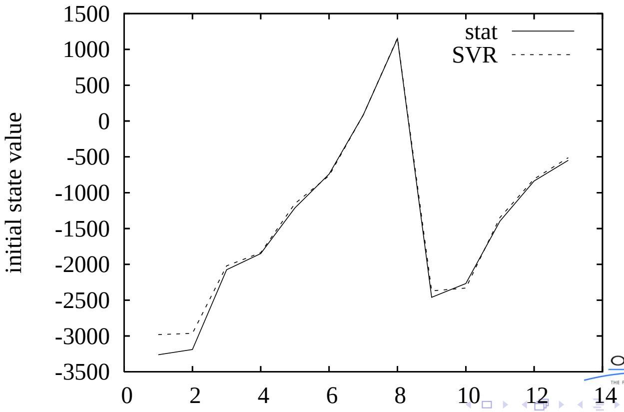
In other words . . .

Which subset of states for API ?

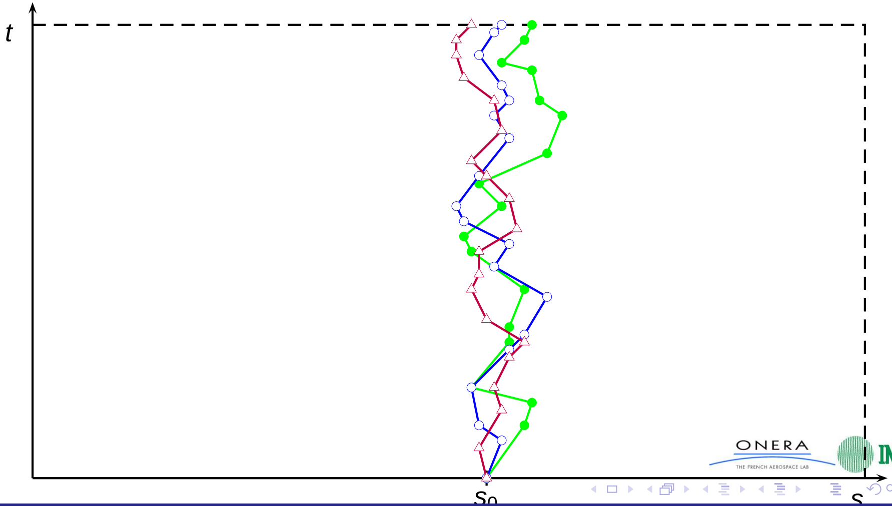
The ones visited by policy simulation !

First results

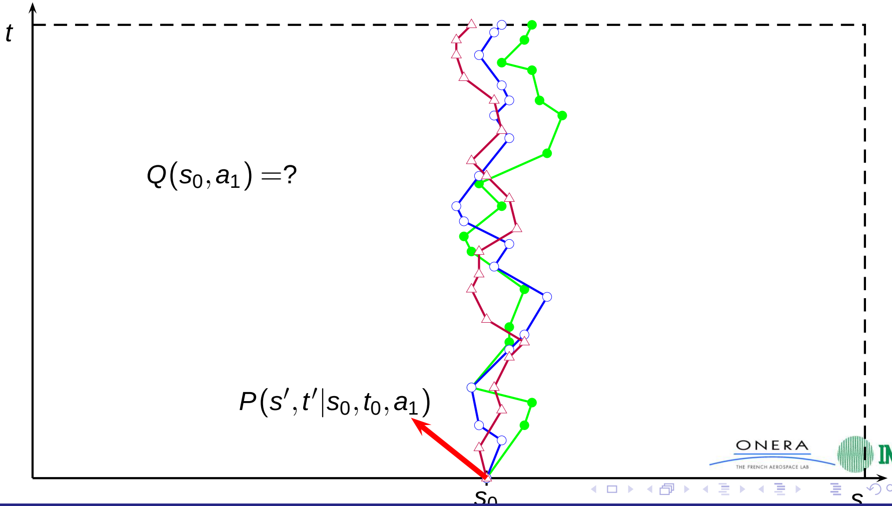
Initial version of online-ATPI with SVR.
Initial policy sets trains to run all day long.



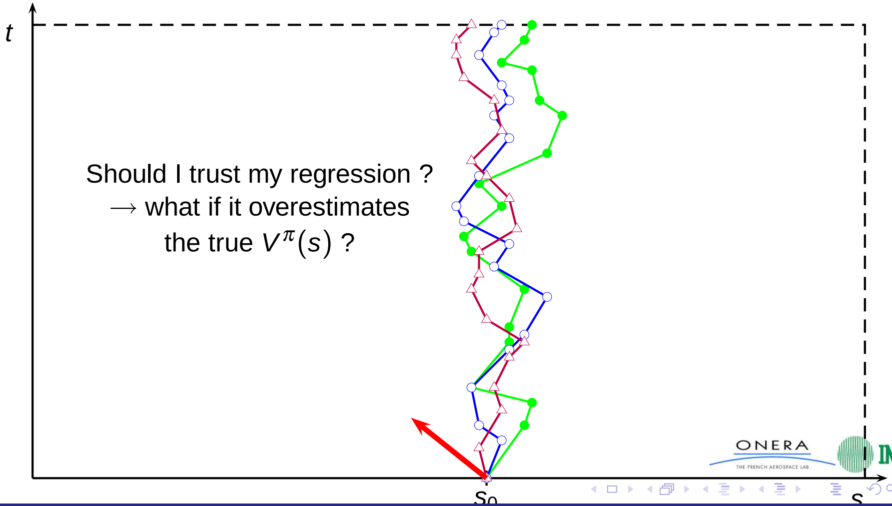
Is there anybody out there ?



Is there anybody out there ?

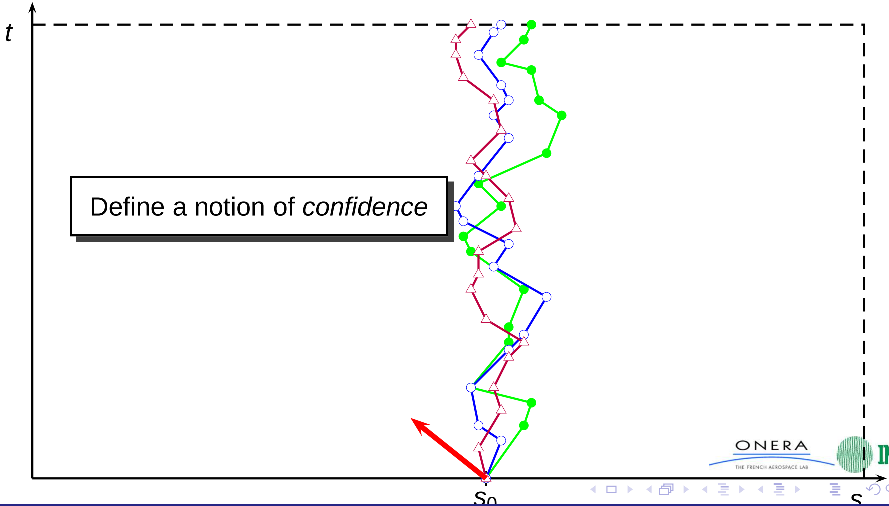


Is there anybody out there ?



Should I trust my regression ?
$\longrightarrow$ what if it overestimates
the true $V^{\pi}(s)$?

Is there anybody out there ?



Introducing confidence

- "confidence" $\Leftrightarrow$ having enough points around s
  $\Leftrightarrow$ approaching the sufficient statistics for $V^{\pi}(s)$
  $\longrightarrow$ approx. measure: pdf of the underlying process.
- What should we do if we are not confident ?
  $\longrightarrow$ generate data – increase the samples' density – simulate
- Storing the policy ?

Same problem for policy storage than for value function:
(*Lagoudakis et al., 03)* RL as Classification.

Full statistical learning problem:
(local incremental) regression (V^{π}), classification (π),
density estimation ($conf$)

Introducing confidence

- “confidence” $\Leftrightarrow$ having enough points around s
  $\Leftrightarrow$ approaching the sufficient statistics for $V^{\pi}(s)$
  $\longrightarrow$ approx. measure: pdf of the underlying process.
- What should we do if we are not confident ?
  $\longrightarrow$ generate data – increase the samples’ density – simulate
- Storing the policy ?

Same problem for policy storage than for value function:
(*Lagoudakis et al., 03)* RL as Classification.

Full statistical learning problem:
(local incremental) regression (V^{π}), classification (π),
density estimation (*conf*)

Introducing confidence

- "confidence" $\Leftrightarrow$ having enough points around s
  $\Leftrightarrow$ approaching the sufficient statistics for $V^{\pi}(s)$
  $\longrightarrow$ approx. measure: pdf of the underlying process.
- What should we do if we are not confident ?
  $\longrightarrow$ generate data – increase the samples' density – simulate
- Storing the policy ?

Same problem for policy storage than for value function:
(*Lagoudakis et al., 03)* RL as Classification.

Full statistical learning problem:
(local incremental) regression (V^{π}), classification (π),
density estimation (*conf*)

Introducing confidence

- “confidence” $\Leftrightarrow$ having enough points around s
  $\Leftrightarrow$ approaching the sufficient statistics for $V^\pi(s)$
  $\longrightarrow$ approx. measure: pdf of the underlying process.
- What should we do if we are not confident ?
  $\longrightarrow$ generate data – increase the samples’ density – simulate
- Storing the policy ?

Same problem for policy storage than for value function:
(*Lagoudakis et al., 03)* RL as Classification.

Full statistical learning problem:
(local incremental) regression (V^π), classification (π),
density estimation (*conf*)

The bigger picture

simulation-based API

$samples \leftarrow \emptyset$
for $i = 1$ to N_{sim} **do**
  while $t < horizon$ **do**
    estimate Q-values
    $s' \leftarrow$ apply best action
    store (s, a, r, s') in *samples*
  end while
end for
$train \tilde{V}^{\pi}(samples)$
$train \tilde{\pi}(samples)$

The bigger picture

estimate $Q(s,a)$

$\tilde{Q}(s,a) \leftarrow 0$
for $i = 1$ to N_a **do**
  $(r,s') \leftarrow$ pick next state
  if $confidence(s') = true$ **then**
    $\tilde{Q}(s,a) \leftarrow \tilde{Q}(s,a) + \frac{r+\tilde{V}^{\pi}(s')}{N_a}$
  else
    $data = simulate(\pi, s')$
    $retrain \tilde{V}^{\pi}(data)$
    $\tilde{Q}(s,a) \leftarrow \tilde{Q}(s,a) + \frac{r+\tilde{V}^{\pi}(s')}{N_a}$
  end if
end for
return $\tilde{Q}(s,a)$

Conclusion

GSMDP Modeling of large scale temporal problems of decision under uncertainty + introduction of a new LSPI-like method, bringing together results from:

- discrete events simulation
- approximate policy iteration
- statistical learning

API A general method inside API

- partial and incremental state space exploration guided by simulation / local policy improvement
- API as statistical learning

GiSMoP C++ library

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

Perspectives

Ongoing work:

- GiSMoP is still under development
- benchmark analysis (especially variance in V^{π})
- interest of regression vs. brute force rollout is still unclear

This work can benefit from:

- Better tuning of regression / classification / density estimation techniques (currently: LWPR / MC-SVM / OC-SVM)
- Non-arbitrary stopping bounds for sampling
- Error bounds
- ...

Thank you for your attention !