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

Emmanuel Rachelson

Patrick Fabiani

Frédérick Garcia

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.
→ planning to coordinate with an uncertain and unstationary environment.
Planning under uncertainty with time dependency.
→ planning to coordinate with an uncertain and unstationary environment.

Should we open more lines?
Planning under uncertainty with time dependency.
→ planning to coordinate with an uncertain and unstationnary environment.

Airplanes taxiing management
Planning under uncertainty with time dependency.
→ planning to coordinate with an uncertain and unstationary environment.

Onboard planning for coordination
Planning under uncertainty with time dependency.
→ planning to coordinate with an uncertain and unstationary environment.

Adding or removing trains?
Subway problem: toy example

Some figures

4 trains, 6 stations
→ 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 controllable 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 controllable dynamics
Temporal Markov Problems: motivation and modeling

Solving large scale GSMDP: ATP

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)
Temporal Markov Problems: motivation and modeling

Solving large scale GSMDP: ATP

Conclusion

GSMDP, (Younes et al., 04)

GSMP, (Glynn, 89)

Several semi-Markov processes affecting the same state space

→ \langle S, E, A, P, F, r \rangle

One process conditionned by the choice of the action undertaken
GSMDP, (Younes et al., 04)

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

\[ \langle S, E, A, P, F, r \rangle \]

One process conditionned by the choice of the action undertaken

\[ P(s' | s_1, e_4) \]
\[ P(s' | s_2, a) \]
Controlling GSMDP

non-Markov behaviour!

→ 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.
Controlling GSMDP

non-Markov behaviour!

→ 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.
Controlling GSMDP

non-Markov behaviour!

→ 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 $V^\pi$
  - 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.
→ (Monte-Carlo) simulation-based policy evaluation.

Statistical learning

Simulating the policy
⇔ Drawing a set of trajectories
⇔ 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.

→ (Monte-Carlo) simulation-based policy evaluation.

Statistical learning

Simulating the policy

↔ Drawing a set of trajectories

↔ 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)
Temporal Markov Problems: motivation and modeling

Solving large scale GSMDP: ATPI

Conclusion

Basic ideas

Approximate Policy Iteration

Approximate evaluation: $V^{\pi_n}$

One-step improvement: $\pi_{n+1}$
Approximate Policy Iteration

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

→ “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 ?
→ “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 ?

→ “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?

\[ Q(s_0, a_1) = ? \]

\[ P(s', t' | s_0, t_0, a_1) \]
Is there anybody out there?

Should I trust my regression?
→ what if it overestimates the true $V^\pi(s)$?
Is there anybody out there?

Define a notion of *confidence*
Introducing confidence

“confidence” ⇔ having enough points around $s$
⇔ approaching the sufficient statistics for $V^\pi(s)$
→ approx. measure: pdf of the underlying process.

What should we do if we are not confident?
→ 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 ($V^\pi$), classification ($\pi$),
density estimation ($conf$)
Introducing confidence

• “confidence” ⇔ having enough points around $s$
  ⇔ approaching the sufficient statistics for $V^\pi(s)$
    → approx. measure: pdf of the underlying process.
• What should we do if we are not confident?
  → 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 ($V^\pi$), classification ($\pi$),
density estimation ($\text{conf}$)
Introducing confidence

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

Same problem for policy storage than for value function:

(\( \text{Lagoudakis et al., 03} \)) RL as Classification.

Full statistical learning problem:

(local incremental) regression (\( V^\pi \)), classification (\( \pi \)),
density estimation (\( \text{conf} \))
Introducing confidence

- “confidence” ⇔ having enough points around $s$
  ⇔ approaching the sufficient statistics for $V^\pi(s)$
  → approx. measure: pdf of the underlying process.
- What should we do if we are not confident?
  → 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 ($V^\pi$), classification ($\pi$),
density estimation ($\text{conf}$)
The bigger picture

**simulation-based API**

```plaintext
samples ← ∅
for i = 1 to N_{sim} do
    while t < horizon do
        estimate Q-values
        s' ← 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

\[
\textbf{estimate } Q(s, a) \\
\tilde{Q}(s, a) \leftarrow 0 \\
\textbf{for } i = 1 \text{ to } N_a \textbf{ do } \\
\quad (r, s') \leftarrow \text{pick next state} \\
\quad \textbf{if } \text{confidence}(s') = \text{true} \textbf{ then } \\
\quad \quad \tilde{Q}(s, a) \leftarrow \tilde{Q}(s, a) + \frac{r + \tilde{V}\pi(s')}{N_a} \\
\quad \textbf{else} \\
\quad \quad data = \text{simulate}(\pi, s') \\
\quad \quad \text{retrain} \tilde{V}\pi(data) \\
\quad \quad \tilde{Q}(s, a) \leftarrow \tilde{Q}(s, a) + \frac{r + \tilde{V}\pi(s')}{N_a} \\
\quad \textbf{end if} \\
\textbf{end for} \\
\textbf{return } \tilde{Q}(s, a)
\]
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](http://emmanuel.rachelson.free.fr/fr/gismop.html)
Perspectives

Ongoing work:

- GiSMoP is still under development
- benchmark analysis (especially variance in $V^\pi$)
- 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!