Non-Stationary Markov Decision Processes a Worst-Case Approach using Model-Based Reinforcement Learning

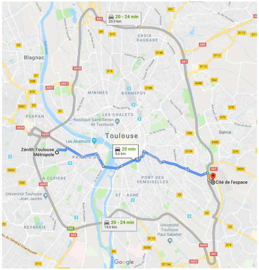
Erwan Lecarpentier
Université de Toulouse
ONERA

Emmanuel Rachelson
Université de Toulouse
ISAE-SUPAERO

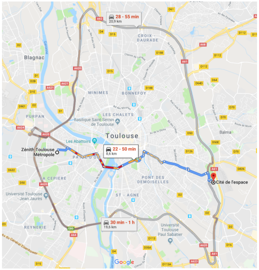
erwan.lecarpentier@isae-supaero.fr

July 3, 2019

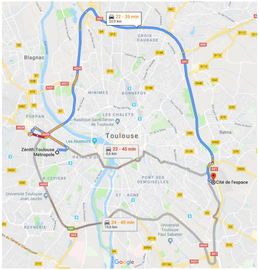
Non-Stationary environment



t_0



t_1



t_2

MDP

Definition 1

A **Markov Decision Process (MDP)** is defined by a 4-tuple $\{\mathcal{S}, \mathcal{A}, r, p\}$ where,

- $\mathcal{S}$ is a state space;
- $\mathcal{A}$ is an action space;
- $r(s, a, s')$ is a reward function;
- $p(s' \mid s, a)$ is the transition probability.

NSMDP

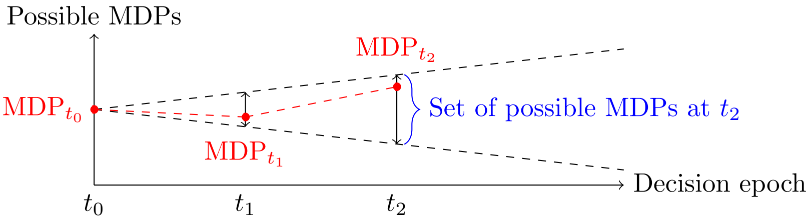
Definition 2

A **Non Stationary Markov Decision Process (NSMDP)** is defined by a 5-tuple $\{\mathcal{S}, \mathcal{T}, \mathcal{A}, r, p\}$ where,

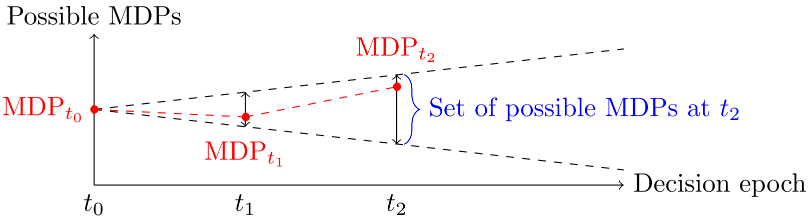
- $\mathcal{S}$ is a state space;
- $\mathcal{T} = \{1, 2, \ldots, N\}$ set of decision epochs, $N \leq +\infty$;
- $\mathcal{A}$ is an action space;
- $r_t(s, a, s')$ is a reward function;
- $p_t(s' \mid s, a)$ is the transition probability.

Two hypotheses

Hypothesis 1: bounded evolution



Hypothesis 1: bounded evolution

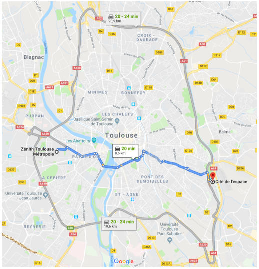


Definition 3

An (L_p, L_r)-**LC-NSMDP** is an NSMDP whose transition and reward functions are respectively L_p-LC and L_r-LC w.r.t. time, i.e., $\forall (t, \hat{t}, s, s', a) \in \mathcal{T}^2 \times \mathcal{S}^2 \times \mathcal{A}$,

$$\begin{cases} W_1(p_t(\cdot \mid s, a), p_{\hat{t}}(\cdot \mid s, a)) & \leq L_p |t - \hat{t}| \\ |r_t(s, a, s') - r_{\hat{t}}(s, a, s')| & \leq L_r |t - \hat{t}|. \end{cases}$$

Hypothesis 2: snapshot

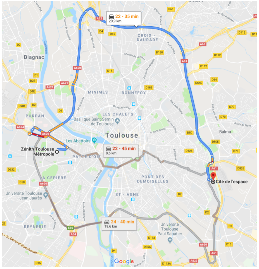


t_0

Known

t_1

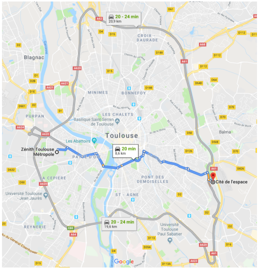
???



t_2

??!

Hypothesis 2: snapshot

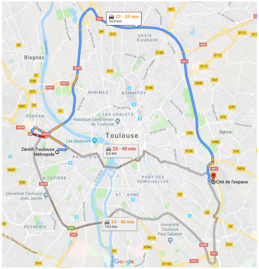


t_0

Known

t_1

???



t_2

??!

Definition 4

The snapshot of an NSMDP $\{\mathcal{S}, \mathcal{T}, \mathcal{A}, (p_t)_{t\in\mathcal{T}}, (r_t)_{t\in\mathcal{T}}\}$ at decision epoch t_0, denoted by MDP_{t_0}, is the stationary MDP defined by the 4-tuple $\{\mathcal{S}, \mathcal{A}, p_{t_0}, r_{t_0}\}$.

So what?

Given a snapshot + the fact that the environment has a bounded evolution, can we find a **robust** plan to **any** evolution of the NSMDP?

Risk Averse Tree-Search

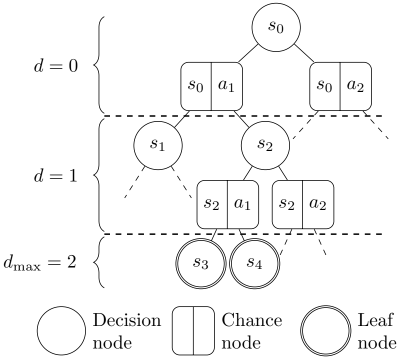


Figure: Tree structure for $\mathcal{A} = \{a_1, a_2\}$ and $d_{\max} = 2$

Risk Averse Tree-Search

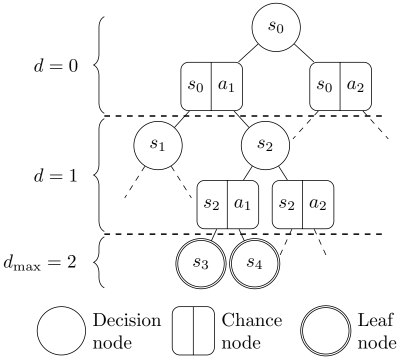


Figure: Tree structure for $\mathcal{A} = \{a_1, a_2\}$ and $d_{\max} = 2$

Decision node value:

$$V(\nu^{s,t}) = \max_{a \in \mathcal{A}} V(\nu^{s,t,a})$$

Chance node value:

$$V(\nu^{s,t,a}) = \min_{(p,R) \in \Delta_{t_0}^{t}} \Big(R(s,a) + \gamma \mathbb{E}_{s' \sim p} V(\nu^{s',t+1})\Big)$$

RATS

Algorithm 1 RATS algorithm

RATS (s_0, t_0, maxDepth)
$\nu_0 = \text{rootNode}(s_0, t_0)$
Minimax(ν_0, maxDepth)
$\nu^* = \arg\max_{\nu \textbf{ in } \nu_0.\text{children}} \nu.\text{value}$
return ν^*.action

Minimax (ν, maxDepth)
if ν is DecisionNode **then**
  if ν.state is terminal **or** $\nu.\text{depth} = \text{maxDepth}$ **then**
    return $\nu.\text{value} = \text{heuristicValue}(\nu.\text{state})$
  else
    return $\nu.\text{value} = \max_{\nu' \in \nu.\text{children}} \text{Minimax}(\nu', \text{maxDepth})$
else
  return $\nu.\text{value} = \min_{(p,R) \in \Delta_{t_0}^{t}} R(\nu) + \gamma \sum_{\nu' \in \nu.\text{children}} p(\nu' \mid \nu) \text{Minimax}(\nu', \text{maxDepth})$

Experiments

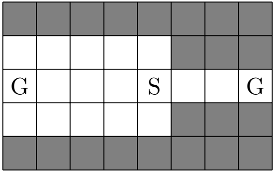


Figure: Non-Stationary bridge environment

- $\epsilon = 0$ left cells get slippery;
- $\epsilon = 1$ right cells get slippery;
- $\epsilon \in (0, 1)$ linear balance between extreme cases.

Experiments

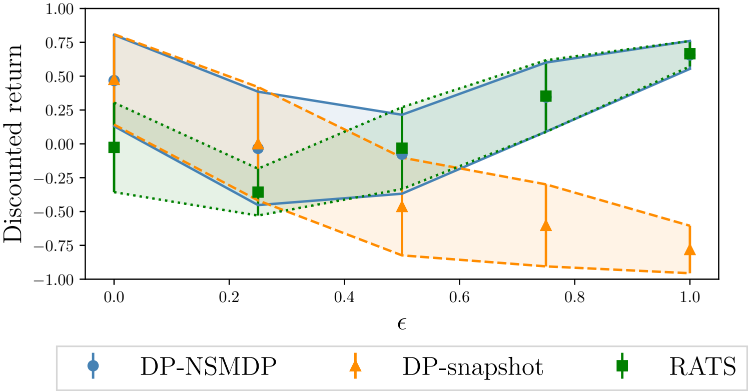


Figure: Discounted return for different values of ϵ.

Experiments

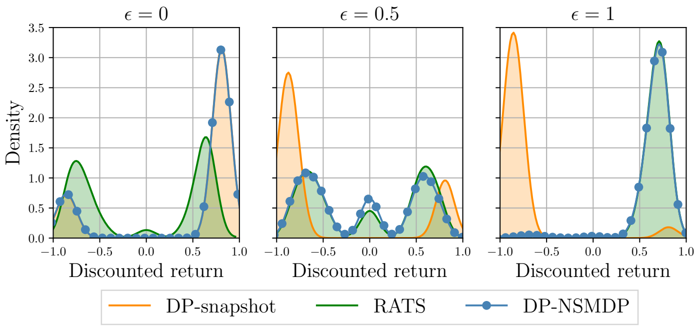


Figure: Discounted return distributions for different values of ϵ.

Supplementary material

Heuristic function

Property 1

Upper bound on the propagated heuristic error within RATS. *Consider an agent executing Algorithm 1 at s_0, t_0 with a heuristic function H. We note $\mathcal{L}$ the set of all leaf nodes. Suppose that the heuristic error is uniformly bounded, i.e.*

$$\exists \delta > 0, \forall \nu^{s,t} \in \mathcal{L}, |H(s) - \hat{V}^*_{t_0,t}(s)| \leq \delta.$$

Then we have for every decision and chance nodes $\nu^{s,t}$ and $\nu^{s,t,a}$, at any depth $d \in [0, d_{\max}]$:

$$|V(\nu^{s,t}) - \hat{V}^*_{t_0,t}(s)| \leq \gamma^{(d_{\max}-d)}\delta$$
$$|V(\nu^{s,t,a}) - \hat{Q}^*_{t_0,t}(s,a)| \leq \gamma^{(d_{\max}-d)}\delta.$$

$$H_1(s) = 0$$

Heuristic function

Property 2

Bounds on the snapshots values. *Let* $s \in \mathcal{S}$, π *a stationary policy,* MDP_{t_0} *and* MDP_t *two snapshot MDPs,* $t, t_0 \in \mathcal{T}^2$ *be. We note* $V^{\pi}_{MDP_i}(s)$ *the value of* s *within* MDP_i *following* π. *Then,*

$$|V^{\pi}_{MDP_{t_0}}(s) - V^{\pi}_{MDP_t}(s)| \leq |t - t_0| \frac{L_r + L_p}{1 - \gamma}.$$

$$H_2(s) = \widehat{V}^{\pi}_{\mathrm{MDP}_{t_0}}(s) - |t - t_0| \frac{L_r + L_p}{1 - \gamma}$$