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

Erwan Lecarpentier
Université de Toulouse
ONERA - The French Aerospace Lab
erwan.lecarpentier@isae-supaero.fr

Emmanuel Rachelson
Université de Toulouse
ISAE-SUPAERO
emmanuel.rachelson@isae-supaero.fr

Abstract
We study MDPs evolving over time (Definition 1) and consider planning in this setting. We make two hypotheses:
1. Continuous, bounded, evolution (Definition 2);
2. Snapshot model (Definition 3) known at each decision epoch.
Our contribution can be presented in three points:
1. Proposal of a planning method robust to the environment’s evolution;
2. Introduction of a zero-shot model-based algorithm, Risk-Averse Tree-Search (RATS), computing the best-worst-case action;
3. Illustration of the benefits of the approach in experiments.

Non-Stationary Markov Decision Processes

Definition (1)
An NSMDP is an MDP whose transition and reward functions depend on the decision epoch. It is defined by a 5-tuple \( \langle S, T, A, (p_t)_{t \in T}, (r_t)_{t \in T} \rangle \) where \( S \) is a state space; \( T \equiv \{1, 2, \ldots, N\} \) is the set of decision epochs, \( N < +\infty \); \( A \) is an action space; \( p_t(s' | s, a) \) is the probability of reaching state \( s' \) with action \( a \) at decision epoch \( t \) in state \( s \); \( r_t(s, a, s') \) is the reward associated to the transition from \( s \) to \( s' \) with action \( a \) at decision epoch \( t \).

Definition (2)
An \( (L_0, L_1)-LC\text{-NSMDP} \) is an NSMDP whose transition and reward functions are respectively \( L_0 \)-LC and \( L_1 \)-LC w.r.t. time, i.e.,
\[
\forall (t, \hat{t}, s, s', a) \in \mathcal{T}^2 \times S^2 \times A, \begin{cases}
W(t)p_t(s | s, a), p_t(s | s, a) & \leq L_0|t - \hat{t}| \\
|r_t(s, a, s')| - r_t(s, a, s') & \leq L_1|t - \hat{t}|.
\end{cases}
\]

Risk-Averse Tree-Search algorithm

Definition (3)
The snapshot of an NSMDP \( \{S, T, A, (p_t)_{t \in T}, (r_t)_{t \in T}\} \) at decision epoch \( t_0 \) denoted by MDP\(_{s0}\) is the stationary MDP defined by the 4-tuple \( \{S, A, p_{s0}, r_{s0}\} \) where \( p_{s0}(s' | s, a) \) and \( r_{s0}(s, a, s') \) are the transition and reward functions of the NSMDP at \( t_0 \).

Proposition (1)
Set of admissible snapshot models. Consider an \( (L_0, L_1)-LC\text{-NSMDP} \), \( s, t, a \in S \times T \times A \). The transition and expected reward functions \( (p_t, R_t) \) of the snapshot MDP, respect
\[
(p_t, R_t) \in \Delta_t := B_{W_t}(p_{t-1} | s, a, L_t) \times B_{R_t}(R_{t-1} | s, a, L_t + L_1)
\]
where \( B_{W_t}(c, r) \) is the ball of centre \( c \), defined with metric \( d \) and radius \( r \).

Algorithm 1: RATS algorithm

RATS \( (s_0, b_0, \text{maxDepth}) \)
\( \nu_0 := \text{rootNode}(s_0, b_0) \)
\( \text{Minimax}(\nu_0) \)
\( \nu' = \arg \max_{\nu} \text{value} \in \nu, \text{children} \nu', \text{value} \)
\( \text{return} \ \nu', \text{action} \)

Minimax \( (\nu, \text{maxDepth}) \)
if \( \nu \) is DecisionNode then
if \( \nu, \text{state} \) is terminal or \( \nu, \text{depth} = \text{maxDepth} \) then
\( \text{return} \ \nu, \text{value} = \text{heuristicValue}(\nu, \text{state}) \)
else
\( \text{return} \ \nu, \text{value} = \max_{\nu' \in \nu, \text{children}} \text{Minimax}(\nu', \text{maxDepth}) \)
else
\( \text{return} \ \nu, \text{value} = \min_{(p, R) \in \Delta_t} R(\nu) + \gamma \sum_{\nu' \in \nu, \text{children}} p(\nu' | \nu) \text{Minimax}(\nu', \text{maxDepth}) \)

Experiments
The value of \( \epsilon \in [0, 1] \) defines different possible evolutions:
- \( \epsilon = 0 \) left cells are slippery;
- \( \epsilon = 1 \) right cells are slippery;
- \( \epsilon \in (0, 1) \) linear balance between extreme cases.

Figure: Discounted return vs \( \epsilon \), 50% of standard deviation.

Figure: Discounted return distributions \( \epsilon \in (0.05, 1) \).