Monte Carlo Tree Search guided by Symbolic Advice for MDPs

Damien Busatto-Gaston, Debraj Chakraborty and Jean-Francois Raskin

Université Libre de Bruxelles

September 16, 2020  HIGHLIGHTS 2020
Markov Decision Process

Path of length 2: $s_0 \xrightarrow{a_1} s_1 \xrightarrow{a_3} s_2$
Markov Decision Process

Path of length 2: \( s_0 \xrightarrow{a_1 \frac{2}{3}} s_1 \xrightarrow{a_3 \frac{1}{2}} s_2 \)

Finite-horizon total reward (horizon \( H \))

\[
\text{Val}(s_0) = \sup_{\sigma: \text{Paths}} \mathbb{E}[\text{Reward}(p)]
\]

where \( p \) is a random variable over \( \text{Paths}^H(s_0, \sigma) \)

Link with infinite-horizon average reward for \( H \) large enough
Monte Carlo tree search (MCTS)

- Iterative construction of a sparse tree with value estimates
Monte Carlo tree search (MCTS)

- Iterative construction of a sparse tree with value estimates
- Selection of a new node $\sim$ simulation
Monte Carlo tree search (MCTS)

- Iterative construction of a sparse tree with value estimates
- Selection of a new node $\leadsto$ simulation $\leadsto$ update of the estimates
Monte Carlo tree search (MCTS)

With UCT (Kocsis & Szepesvári, 2006) as the selection strategy:

- After a given number of iterations $n$, MCTS outputs the best action
- The probability of choosing a suboptimal action converges to zero
- $v_i$ converges to the real value of $a_i$ at a speed of $(\log n)/n$
Symbolic advice
Symbolic advice

An advice is a subset of Paths\(^H(s_0)\)

- Defined symbolically as a logical formula \(\varphi\) (reachability or safety property, LTL formula over finite traces, regular expression ...)
Symbolic advice

- An **advice** is a subset of Paths$^H(s_0)$
- Defined symbolically as a logical formula $\varphi$ (reachability or safety property, LTL formula over finite traces, regular expression ...)
Symbolic advice

- An advice is a subset of $\text{Paths}^H(s_0)$
- Defined symbolically as a logical formula $\varphi$ (reachability or safety property, LTL formula over finite traces, regular expression . . . )
Symbolic advice

Strongly enforceable advice: can be enforced by controller if the MDP is seen as a game ~ does not partially prune stochastic transitions
Symbolic advice

Strongly enforceable advice: can be enforced by controller if the MDP is seen as a game does not partially prune stochastic transitions
Symbolic advice

Strongly enforceable advice: can be enforced by controller if the MDP is seen as a game \( \sim \) does not partially prune stochastic transitions
The advice $\psi$ can be encoded as a Boolean Formula.

- Inuctive way of constructing paths that satisfy the strongly enforceable advice $\phi$.
- Weighted sampling.
- Simulation of safe paths according to $\psi$.
- Weighted SAT sampling (Chakraborty, Fremont, Meel, Seshia, & Vardi, 2014).
The advice $\psi$ can be encoded as a Boolean Formula

QBF solver

A first action $a_0$ is compatible with $\varphi$ iff

$$\forall s_1 \exists a_1 \forall s_2 \ldots, s_0 a_0 s_1 a_1 s_2 \ldots \models \psi$$

Inductive way of constructing paths that satisfy the strongly enforceable advice $\varphi$
The advice $\psi$ can be encoded as a Boolean Formula

$$QBF \text{ solver}$$

A first action $a_0$ is compatible with $\varphi$ iff

$$\forall s_1 \exists a_1 \forall s_2 \ldots, s_0 a_0 s_1 a_1 s_2 \ldots \models \psi$$

Inductive way of constructing paths that satisfy the strongly enforceable advice $\varphi$

Weighted sampling

Simulation of safe paths according to $\psi$

Weighted SAT sampling (Chakraborty, Fremont, Meel, Seshia, & Vardi, 2014)
MCTS under advice
Select actions in the unfolding pruned by a selection advice $\varphi$

Simulation is restricted according to a simulation advice $\psi$
MCTS under advice

Convergence properties

With UCT (Kocsis & Szepesvári, 2006) as the selection strategy:

- The probability of choosing a suboptimal action converges to zero
- $v_i$ converges to the real value of $a_i$ at a speed of $(\log n)/n$

The convergence properties are maintained:

- for all simulation advice
- for all selection advice which
Convergence properties

With **UCT** (Kocsis & Szepesvári, 2006) as the selection strategy:

- The probability of choosing a suboptimal action converges to zero
- \( v_i \) converges to the real value of \( a_i \) at a speed of \( (\log n)/n \)

The convergence properties are maintained:

- for all simulation advice
- for all selection advice which . . .
  - are **Strongly enforceable advice**
Convergence properties

With UCT (Kocsis & Szepesvári, 2006) as the selection strategy:

- The probability of choosing a suboptimal action converges to zero
- \( v_i \) converges to the real value of \( a_i \) at a speed of \( (\log n)/n \)

The convergence properties are maintained:

- for all simulation advice
- for all selection advice which
  - are Strongly enforceable advice
  - satisfy an optimality assumption: does not prune all optimal actions
Experimental results
## Experimental results

**Figure:** 9x21 maze, 4 random ghosts

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>% of win</th>
<th>% of loss</th>
<th>% of no result(^1)</th>
<th>% of food eaten</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCTS</td>
<td>17</td>
<td>59</td>
<td>24</td>
<td>67</td>
</tr>
<tr>
<td>MCTS+Selection advice</td>
<td>25</td>
<td>54</td>
<td>21</td>
<td>71</td>
</tr>
<tr>
<td>MCTS+Simulation advice</td>
<td>71</td>
<td>29</td>
<td>0</td>
<td>88</td>
</tr>
<tr>
<td>MCTS+both advice</td>
<td>85</td>
<td>15</td>
<td>0</td>
<td>94</td>
</tr>
<tr>
<td>Human</td>
<td>44</td>
<td>56</td>
<td>0</td>
<td>75</td>
</tr>
</tbody>
</table>

\(^1\)after 300 steps
Future works

- Compiler LTL $\rightarrow$ symbolic advice
- Study interactions with reinforcement learning techniques (and neural networks)
- Weighted advice
Future works

- Compiler LTL → symbolic advice
- Study interactions with reinforcement learning techniques (and neural networks)
- Weighted advice

Thank You