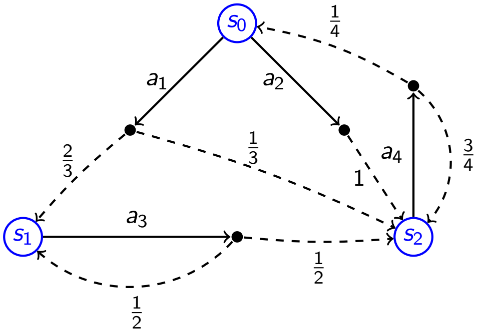
Monte Carlo Tree Search guided by Symbolic Advice for MDPs

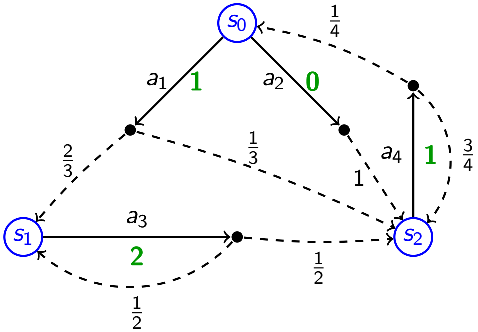
Damien Busatto-Gaston, Debraj Chakraborty
and Jean-Francois Raskin

Université Libre de Bruxelles

September 16, 2020 HIGHLIGHTS 2020

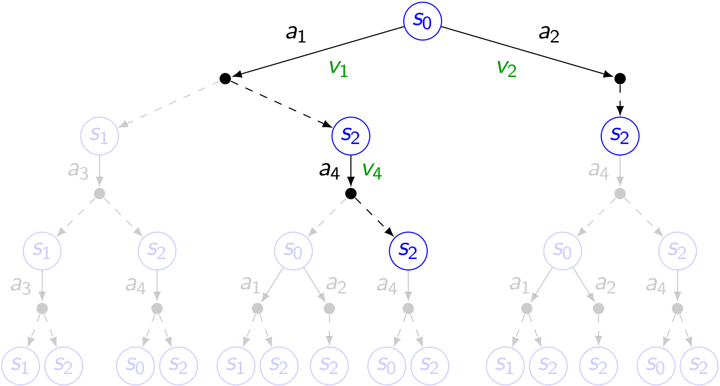


- Path of length 2: $s_0 \xrightarrow{a_1} \overset{\frac{2}{3}}{\dashrightarrow} s_1 \xrightarrow{a_3} \overset{\frac{1}{2}}{\dashrightarrow} s_2$



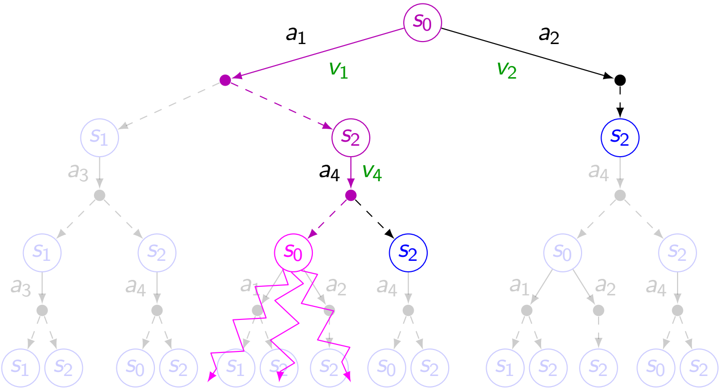
- Path of length 2: $s_0 \xrightarrow{a_1} \overset{\frac{2}{3}}{\dashrightarrow} s_1 \xrightarrow{a_3} \overset{\frac{1}{2}}{\dashrightarrow} s_2$
- Finite-horizon total reward (horizon H)
- $\mathrm{Val}(s_0) = \sup_{\sigma:\mathrm{Paths}\to A} \mathbb{E}\left[\mathrm{Reward}(p)\right]$
  where p is a random variable over $\mathrm{Paths}^H(s_0, \sigma)$
- Link with infinite-horizon average reward for H large enough

Monte Carlo tree search (MCTS)



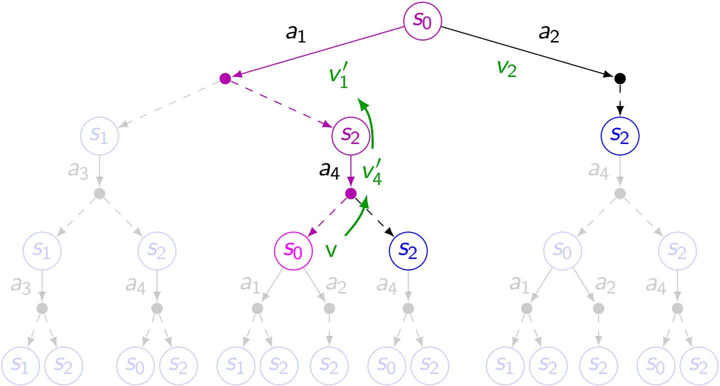
- Iterative construction of a sparse tree with value estimates

Monte Carlo tree search (MCTS)

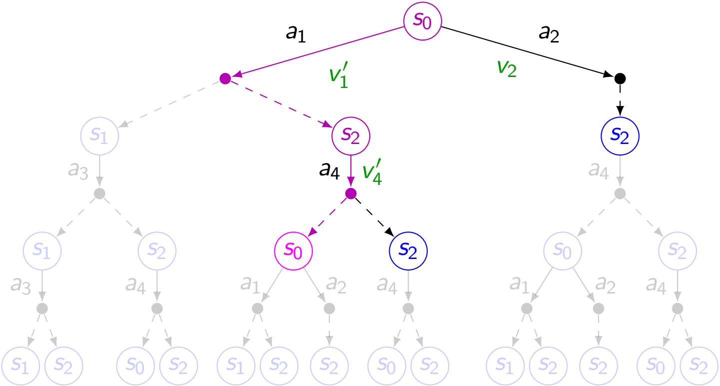


- Iterative construction of a sparse tree with value estimates
- Selection of a new node $\rightsquigarrow$ simulation

Monte Carlo tree search (MCTS)



- Iterative construction of a sparse tree with value estimates
- Selection of a new node $\rightsquigarrow$ simulation $\rightsquigarrow$ update of the estimates

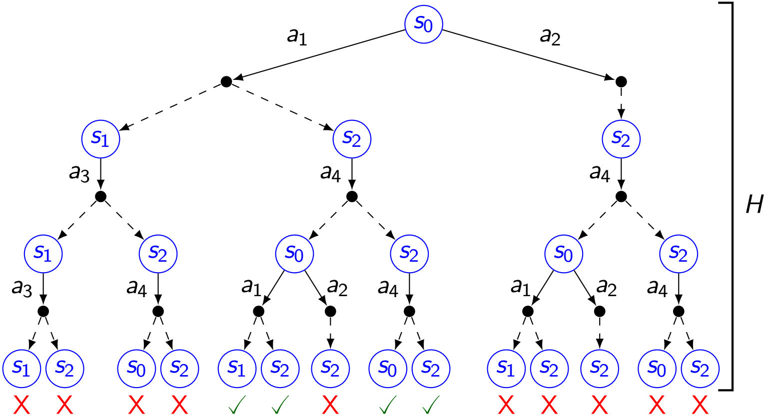


With UCT (Kocsis & Szepesvári, 2006) as the selection strategy:

- 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 $(\log n)/n$

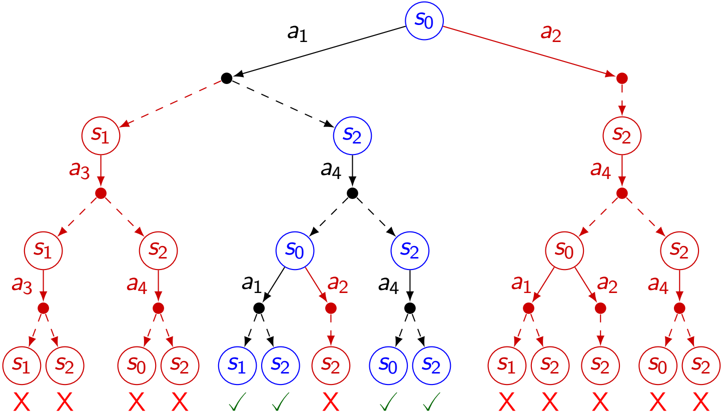
Symbolic advice

Symbolic advice



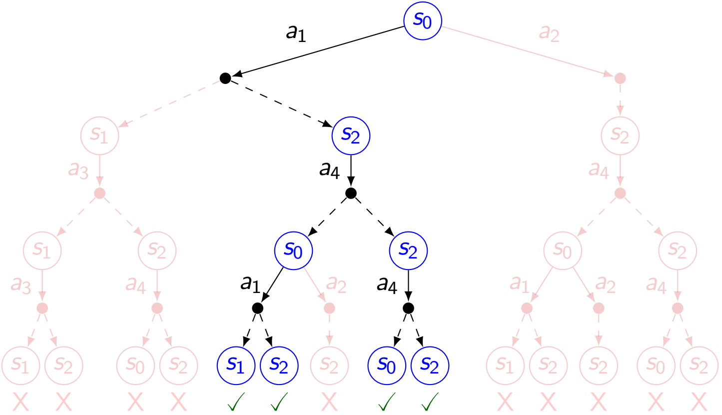
- An advice is a subset of $\mathrm{Paths}^H(s_0)$
- Defined symbolically as a logical formula φ (reachability or safety property, LTL formula over finite traces, regular expression ...)

Symbolic advice

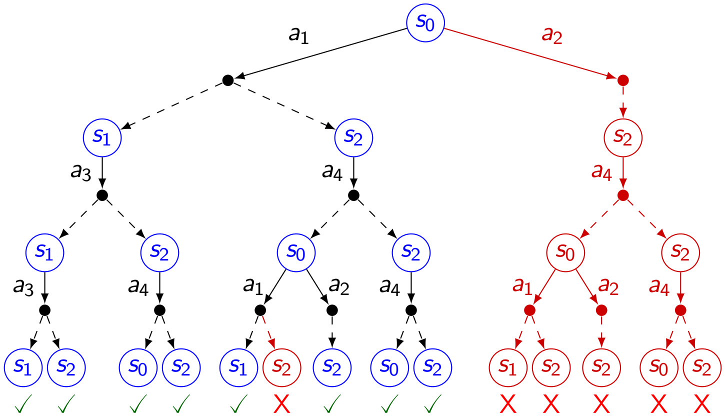


- An advice is a subset of $\mathrm{Paths}^H(s_0)$
- Defined symbolically as a logical formula φ (reachability or safety property, LTL formula over finite traces, regular expression . . .)

Symbolic advice

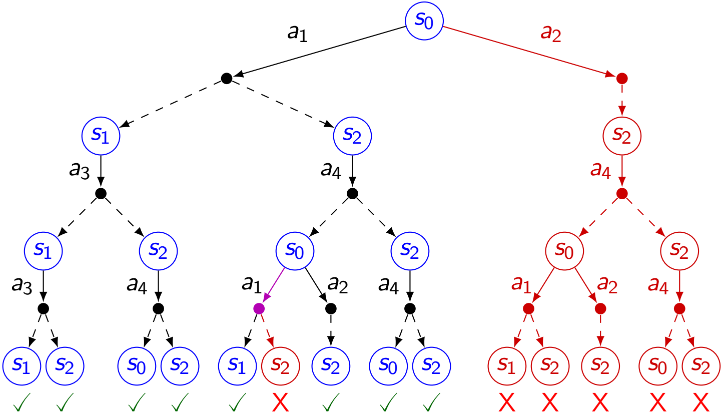


- An advice is a subset of $\text{Paths}^H(s_0)$
- Defined symbolically as a logical formula φ (reachability or safety property, LTL formula over finite traces, regular expression ...)



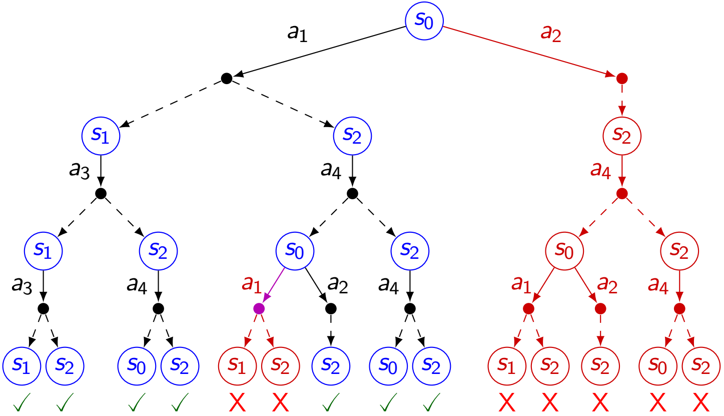
Strongly enforceable advice: can be enforced by controller if the MDP is seen as a game $\rightsquigarrow$ does not partially prune stochastic transitions

Symbolic advice



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

Symbolic advice



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

Boolean Solvers

- The advice ψ can be encoded as a Boolean Formula

- The advice ψ can be encoded as a Boolean Formula

QBF solver

- A first action a_0 is compatible with φ 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 φ

- The advice ψ can be encoded as a Boolean Formula

QBF solver

- A first action a_0 is compatible with φ iff

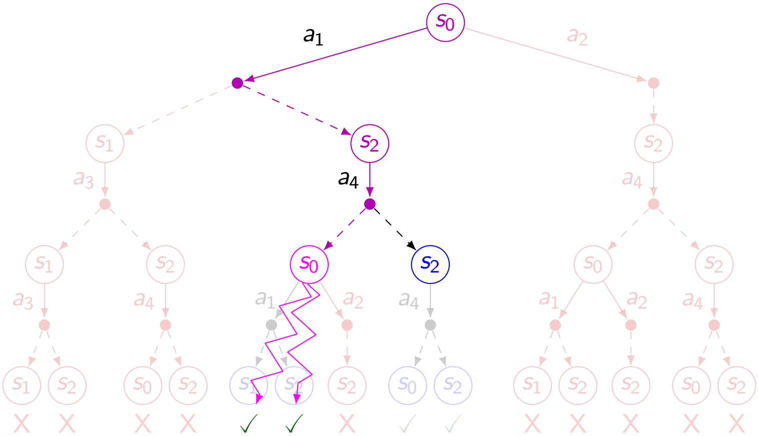
$$\forall s_1 \exists a_1 \forall s_2 \ldots, \; s_0 a_0 s_1 a_1 s_2 \ldots \models \psi$$

- Inductive way of constructing paths that satisfy the strongly enforceable advice φ

Weighted sampling

- Simulation of safe paths according to ψ
- Weighted SAT sampling (Chakraborty, Fremont, Meel, Seshia, & Vardi, 2014)

MCTS under advice



- Select actions in the unfolding pruned by a selection advice φ
- Simulation is restricted according to a simulation advice ψ

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

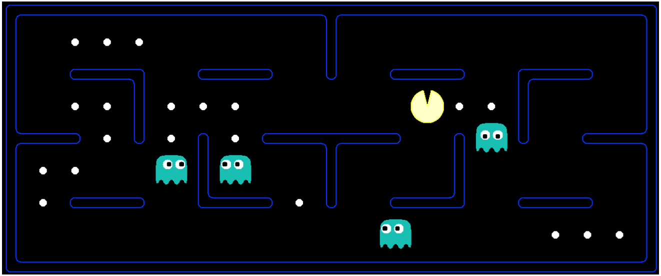


Figure: 9x21 maze, 4 random ghosts

Algorithm	% of win	% of loss	% of no result[1]	% of food eaten
MCTS	17	59	24	67
MCTS+Selection advice	25	54	21	71
MCTS+Simulation advice	71	29	0	88
MCTS+both advice	85	15	0	94
Human	44	56	0	75

[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 $\rightarrow$ symbolic advice
- Study interactions with reinforcement learning techniques (and neural networks)
- Weighted advice

Thank You