Value Iteration for Simple Stochastic Games: Stopping Criterion and Learning Algorithm

E. Kelmedi J. Krämer-Eisentraut J. Křetínský M. Weininger

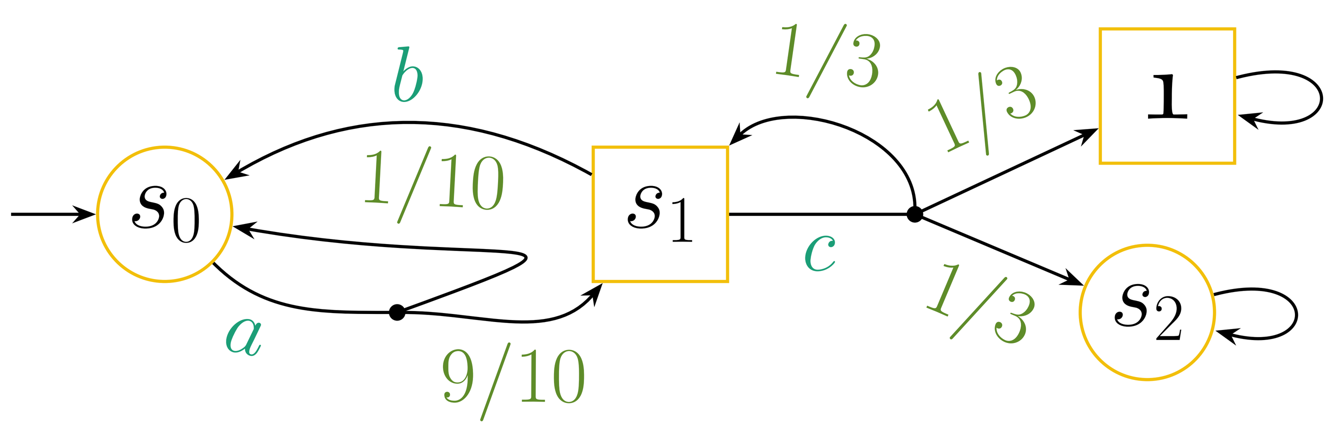
Fakultät für Informatik, Technische Universität München

Our Contributions[3]

- First convergent anytime algorithm with guaranteed precision.
- Learning-based variant often needs only fraction of state space.

Reachability in Simple Stochastic Games (SG)

- States S, actions A and transition probabilities δ.
- States belong to one of two players: Maximizer □ or Minimizer ○
- Value = Probability to reach goal state $\mathtt{1}$ if both play optimally, i.e. $V(s) = \sup_\sigma \inf_\tau \mathbb{P}_s^{\sigma,\tau}(\diamond \mathtt{1}) = \inf_\tau \sup_\sigma \mathbb{P}_s^{\sigma,\tau}(\diamond \mathtt{1})$.
- Compute $V(s_0)$ as well as optimal strategies σ, τ.



Value Iteration (VI)

Bellman update

$$f_{i+1}(s) = \begin{cases} \max_{a \in A} f_i(s,a) & \text{if } s \text{ belongs to } \square \\ \min_{a \in A} f_i(s,a) & \text{if } s \text{ belongs to } \bigcirc \end{cases}$$

where $f_i(s,a) = \Sigma_{s' \in S}\, \delta(s,a,s') \cdot f_i(s')$

- The value V is the least fixpoint of the Bellman equations.
- Applying Bellman updates to under-approximation $L_0(s) = \begin{cases} 1 & \text{if } s = \mathtt{1} \\ 0 & \text{otherwise} \end{cases}$ yields $\lim_{i\to\infty} L_i = V$
  BUT we do not know how close any L_i is to V, i.e. when to stop.
- By applying Bellman updates to an over-approximation $U_0(s) = 1$ we get a guaranteed interval,
  BUT U need not converge to V, but some greater fixpoint.

	Normal			+ Deflating	
Iteration i	$L(s_0)$	$L(s_1)$	$U(s_1)$	$U(s_1)$	$U(s_2)$
0	0	0	1	1	1
1	0	$1/3$	1	1	0
2	$9/30$	$4/9$	1	$2/3$	0
3	$43/100$	$13/27$	1	$5/9$	0

End Components (EC)

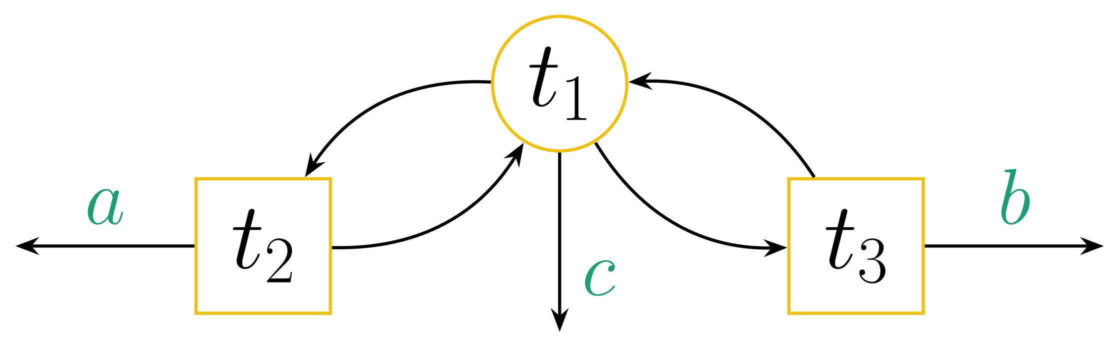
- An EC is a set of states $T \subseteq S$, where under some pair of strategies a play reaching T remains there forever.
- E.g. $T = \{s_0, s_1\}$ in Figure 1 (if s_1 chooses b).

The Cause of Non-Convergence: Simple End Components (SEC)

- An EC is a SEC, if it only uses optimal actions of Minimizer.
- Assigning any $m \in \mathbb{R}$ with $V(\text{bestExit}_\square) \leq m \leq V(\text{bestExit}_\bigcirc)$ to all states in a SEC locally solves the Bellman equations.

- E.g. $\{s_0, s_1\}$ also is a SEC, with $m \in [0.5, 1]$.
- The figure below is parametrized, to show that depending on the values there can be different SECs in an EC.

Minimal value	$V(t_2,a)$	$V(t_3,b)$	$V(t_2,a) \wedge V(t_3,b)$	$V(t_1,c)$
SEC	$\{t_1,t_2\}$	$\{t_1,t_3\}$	$\{t_1,t_2,t_3\}$	$\emptyset$



Deflating SECs

- We "deflate" a SEC by reducing all upper bounds to $U(\text{bestExit}_\square)$.
- **Soundness:** Deflating is sound for any set of states.
- We guess the SECs according to the current L.
- **Correctness:** Since L converges to V, we eventually find and deflate the true SECs.

Relation to MDP algorithms

- In MDP, every EC is a SEC.
- The approach for MDPs[1] works on SECs. As we might only find them in the limit, it does not generalize to SG.
- The learning-based algorithm for MDP[1] is extended by replacing the former EC treatment with deflating.

Implementation

- Implemented both algorithms as an extension of PRISM-games[2].
- The computational overhead for the additional over-approximation often is negligible.

Future Work

Give convergent algorithm with stopping criterion for SG

- with other objectives, e.g. total reward, mean payoff, omega-regular.
- with multi-objective queries.
- in limited information settings.
- based on other learning algorithms.

[1] T. Brázdil, K. Chatterjee, M. Chmelik, V. Forejt, J. Kretínský, M. Z. Kwiatkowska, D. Parker, and M. Ujma. Verification of markov decision processes using learning algorithms. In *ATVA 2014, Proceedings*, pages 98–114, 2014.
[2] T. Chen, V. Forejt, M. Z. Kwiatkowska, D. Parker, and A. Simaitis. Prism-games: A model checker for stochastic multi-player games. In *TACAS 2013, Proceedings*, pages 185–191, 2013.
[3] E. Kelmendi, J. Krämer, J. Kretínský, and M. Weininger. Value iteration for simple stochastic games: Stopping criterion and learning algorithm. In *CAV 2018, Proceedings, Part I*, pages 623–642, 2018.