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:
- 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 \( \delta \).
- States belong to one of two players: Maximizer \( \square \) or Minimizer \( \bigcirc \).
- Value = Probability to reach goal state \( 1 \) if both play optimally, i.e. \( V(s) = \sup_{\pi} \inf_{\tau} P^{\pi,\tau}(s,1) = \inf_{\tau} \sup_{\pi} P^{\pi,\tau}(s,1) \).
- Compute \( V(s) \) as well as optimal strategies \( \sigma, \tau \).

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 \( f_i(s,a) = \sum_{s' \in S} \delta(s,a,s') \cdot f_i(s') \)

- The value \( V \) is the least fixpoint of the Bellman equations.
- Applying Bellman updates to under-approximation \( L_i(s) = \begin{cases} 1 & \text{if } s = 1 \\ 0 & \text{otherwise} \end{cases} \) 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 \( U_i(s) = 1 \) we get a guaranteed interval,
- BUT \( U \) need not converge to \( V \), but some greater fixpoint.

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Normal</th>
<th>Deflating</th>
</tr>
</thead>
<tbody>
<tr>
<td>( L(s_0) )</td>
<td>( L(S) )</td>
<td>( U(s) )</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1/3</td>
</tr>
<tr>
<td>2</td>
<td>9/30</td>
<td>4/9</td>
</tr>
<tr>
<td>3</td>
<td>43/100</td>
<td>13/27</td>
</tr>
</tbody>
</table>

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.

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.

References: