BRTDP for Stochastic Games*

Maximilian Weininger

Technical University of Munich

Abstract. Simple stochastic games can be solved by value iteration (VI), which yields a sequence of under-approximations of the value of the game. This sequence is guaranteed to converge to the value only in the limit. Since no stopping criterion is known, this technique does not provide any guarantees on its results. We provide the first stopping criterion for VI on simple stochastic games. It is achieved by additionally computing a convergent sequence of *over-approximations* of the value, relying on an analysis of the game graph. Consequently, VI becomes an anytime algorithm returning the approximation of the value and the current error bound. As another consequence, we can provide a simulation-based asynchronous VI algorithm, which yields the same guarantees, but without necessarily exploring the whole game graph.

Simple stochastic games (SG) can be solved by multiple algorithms [Con93]. Value iteration (VI) is usually preferred, as it typically is the fastest method [ACD+17]. However, VI may converge only in the limit, and prior to our work there was no known stopping criterion for VI applied to SG. Consequently, there were no guarantees on the results returned in finite time, and they could be arbitrarily imprecise [HM18].

The solution for the special case of Markov decision processes (MDP) was to employ a bounded variant of VI [MLG05,BCC+14] (also called *interval iteration* [HM18]). Here one computes not only an under-approximation, but also an over-approximation of the actual value by iterative computation of the least and greatest fixpoints of the Bellman equations. Since the fixpoints may not coincide (and in fact the greatest fixpoint often results in the trivial bound of 1), additional steps have to be taken. The solution for MDP, namely to modify the underlying graph by collapsing end components, is not applicable for general SG, since there states in an end component can have different values.

Instead, in [KKKW18] we introduced a modified value iteration procedure, where the greatest fixpoint coincides with the value. The key idea is to analyze the game graph and identify special end components, where all states have the same value. By removing a circular dependency in the computation of the over-approximation, the modified value iteration converges. Since the special end components can only be safely identified in the limit, we cannot handle them a priori, but only on-the-fly.

Additionally, we showed how to use simulations and reinforcement learning similar to [MLG05,BCC+14,ACD+17], which sometimes gives speedups of several orders of magnitude. For a more detailed view, we refer the reader to the original paper and the appendix of the technical report available at https://arxiv.org/abs/1804.04901.

* This paper reports on the work published in [KKKW18].

References

[ACD+17] Pranav Ashok, Krishnendu Chatterjee, Przemyslaw Daca, Jan Kretínský, and Tobias Meggendorfer. Value iteration for long-run average reward in markov decision processes. In *CAV*, pages 201–221, 2017.

[BCC+14] Tomás Brázdil, Krishnendu Chatterjee, Martin Chmelik, Vojtech Forejt, Jan Kretínský, Marta Z. Kwiatkowska, David Parker, and Mateusz Ujma. Verification of Markov decision processes using learning algorithms. In *ATVA*, pages 98–114. Springer, 2014.

[Con93] Anne Condon. On algorithms for simple stochastic games. In *Advances in Computational Complexity Theory, volume 13 of DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, pages 51–73. American Mathematical Society, 1993.

[HM18] Serge Haddad and Benjamin Monmege. Interval iteration algorithm for mdps and imdps. *Theor. Comput. Sci.*, 735:111–131, 2018.

[KKKW18] Edon Kelmendi, Julia Krämer, Jan Kretínský, and Maximilian Weininger. Value iteration for simple stochastic games: Stopping criterion and learning algorithm. In *CAV*, 2018.

[MLG05] H. Brendan Mcmahan, Maxim Likhachev, and Geoffrey J. Gordon. Bounded real-time dynamic programming: Rtdp with monotone upper bounds and performance guarantees. In *In ICML05*, pages 569–576, 2005.