Value Iteration for Mean-payoff in Markov Decision Processes

Pranav Ashok[1], Krishnendu Chatterjee[2], Przemysław Daca[2], Jan Křetínský[1], and Tobias Meggendorfer[1]

[1] Technical University of Munich, Germany
[2] IST Austria

Markov decision processes (MDPs) are standard models for probabilistic systems with non-determi-nistic behaviors. The mean-payoff objective provides a mathematically elegant formalism to express performance related properties. The value-iteration (VI) approach provides one of the simplest and most efficient algorithmic approaches for MDPs with other properties (such as reachability objectives). Unfortunately, the straightforward VI approach does not work for the mean-payoff objective in MDPs. In particular, there is no stopping criterion which can give guarantees that the solution obtained through VI is ε-close to the optimal solution.

In our work, the contributions are threefold. (i) We refute a conjecture (presented in [Put14]) related to stopping criteria for mean-payoff objectives in MDPs; (ii) we present two practical algorithms for the mean-payoff objective in MDPs based on VI; and (iii) we present experimental results showing that our approach significantly outperforms the standard approaches on several benchmarks.

A core idea which we exploit is the fact that for infinite horizon objectives like mean-payoff, only rewards which can be obtained in maximal end-components (MECs) of the MDP matter. Keeping this in mind, we present the following two algorithms.

In the first, we show that a combination of local VI in MECs and VI for reachability objectives can provide approximation guarantees. The stopping criterion is known [Put14] for VI in communicating MDPs, the class of MDPs in which for every pair of two states s_i and s_j, there exists a deterministic strategy under which s_i is reachable from s_j. We first identify different MECs in the MDP and then use the fact that a MEC can be identified with a communicating MDP, to run VI on it until the desired precision is achieved. Next, we collapse these MECs into their representative states and reduce the problem of solving the mean-payoff objective to a problem of computing the reachability objective on a transformed MDP. The algorithm presented in [BCC$^+$14], which is a version of asynchronous value iteration using sampling, allows us to compute the reachability objective with the desired guarantees.

In the second, we present an anytime algorithm based on the bounded real-time dynamic programming (BRTDP) approach [MLG05]. In the BRTDP approach, paths are repeatedly sampled in the MDP directed by a heuristc, and VI is applied only for the states on these paths. Moreover, upper and lower bounds are maintained for each state and the VI operator is applied on them

separately. The difference between the upper and lower bounds for every state is a natural measure of error. In addition to this standard approach, we equip our algorithm to detect when a path gets stuck in some end-component, in the lines of [BCC+14]. The end component is then collapsed into a representative state in a collapsed MDP, but unlike earlier, we do not wait for the stopping criterion to be satisfied while running VI on the end-component. The core idea of this approach is that if the probability of reaching a state (or a MEC representative state) is quite low, then we might not need to explore it (or obtain an ε-precise solution in the MEC). Our algorithm is able to detect when the solution in a MEC needs to be refined. It continues alternating between propagating bounds in the collapsed MDP and refining the values in the MECs through running VI locally. Furthermore, the biggest advantage of this algorithm, as in BRTDP, is that not all states need to be explored in order to get approximation guarantees.

Finally, we present the results of our benchmark against MultiGain [BCFK15], the only tool we are aware of, which can solve mean-payoff objectives with guarantees. Our results show that depending on the underlying structure of the MDP, our approaches behave comparable in performance. But on some large models, the second approach is able to obtain a solution within seconds whereas standard methods runs out of memory. In every non-trivial example, both our methods significantly outperform the Linear Programming based approach of MultiGain.

This work has been presented at CAV 2017.

References

[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.

[BCFK15] Tomás Brázdil, Krishnendu Chatterjee, Vojtech Forejt, and Antonín Kucera. MultiGain: A controller synthesis tool for MDPs with multiple mean-payoff objectives. In *TACAS*, pages 181–187, 2015.

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

[Put14] Martin L. Puterman. *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons, 2014.