Learning-Based Mean-Payoff Optimization in Unknown Markov Decision Processes under Omega-Regular Constraints*

Jan Křetínský[1], Guillermo A. Pérez[2], and Jean-François Raskin[3]

[1] Technische Universität München, Munich, Germany
[2] Universiteit Antwerpen, Antwerp, Belgium
[3] Université libre de Bruxelles, Brussels, Belgium

Abstract. We formalize the problem of maximizing the mean-payoff value with high probability while satisfying a parity objective in a Markov decision process (MDP) with unknown probabilistic transition function and unknown reward function. Assuming the support of the unknown transition function and a lower bound on the minimal transition probability are known in advance, we show tight bounds for achievable combinations of types of strategies (finite or infinite memory) and types of guarantees (sure, almost sure, probably approximately correct).

Reactive synthesis and online reinforcement learning. Reactive systems are systems that maintain a continuous interaction with the environment in which they operate. When designing such systems, we usually face two partially conflicting objectives. First, to ensure a safe execution, we want some basic and critical properties to be enforced by the system no matter how the environment behaves. Second, we want the reactive system to be as efficient as possible given the actual observed behaviour of the environment in which the system is executed. As an illustration, let us consider a robot that needs to explore an unknown environment as efficiently as possible while avoiding any collision. While operating at low speed makes it easier to avoid collisions, it will impair its ability to explore the environment quickly even if the environment is clear of other objects.

There has been, in the past, a large research effort to define mathematical models and algorithms in order to address the two objectives above, but in isolation only. To synthesize safe control strategies, two-player zero-sum games with omega-regular objectives have been proposed, see e.g. [1]. Reinforcement-learning (RL, for short) algorithms for partially-specified Markov decision processes (MDPs) have been proposed, see e.g. [5]) to learn strategies that reach (near-)optimal performance in the actual environment in which the system is executed. Here we want to answer the following question: *How efficient can online-learning techniques be if only correct executions, i.e. executions that satisfy a specified omega-regular objective, are explored during execution?* So, we want to understand how to combine synthesis and RL to construct systems that are safe, yet, at the same time, can adapt their behaviour according to the actual environment in which they execute.

* This abstract is a shortened version of the Concur'18 paper [4].

Problem statement. In order to answer in a precise way the question above, we consider a model halfway between the fully-unknown models considered in RL and the full-known models used in verification. To be precise, we consider as input an MDP with rewards whose transition probabilities are not known and whose rewards are discovered on the fly. That is, the input is the support of the unknown transition function of the MDP. This is natural from the point of view of verification since: we may be working with an underspecified system, its qualitative behaviour may have already been observed, or we may not trust all given probability values. As optimization objective on this MDP, we consider the mean-payoff function, and to capture the sure omega-regular constraint we use a parity objective.

Contributions. Given a lower bound $\pi_{\min}$ on the minimal transition probability, we show that, in partially-specified MDPs consisting of a single end component (EC), two combinations of guarantees on the parity and mean-payoff objectives can be achieved:

1. For all ε and γ, we show how to construct a finite-memory strategy which almost-surely satisfies the parity objective and which achieves an ε-optimal mean payoff with probability at least $1 - \gamma$.
2. For all ε and γ, we show how to construct an infinite-memory strategy which satisfies the parity objective surely and which achieves an ε-optimal mean payoff with probability at least $1 - \gamma$.

We also extend our results to MDPs consisting of more than one EC in a natural way and study special cases that allow for improved optimality results as in the case of good ECs. Finally, we show that there are partially-specified MDPs for which stronger combinations of the guarantees cannot be ensured. Our usage of $\pi_{\min}$ follows [2,3] where it is argued that it is necessary for the statistical analysis of unbounded-horizon properties and realistic in many scenarios.

Example: Almost-sure constraints. Consider the MDP on the right-hand side of Fig. 1 for which we know the support of the transition function but not the probabilities x and y (for simplicity the rewards are assumed to be known). First, note that while there is no surely winning strategy for the parity objective in this MDP, playing action a forever in q_0 guarantees to visit state q_3 infinitely many times with probability one, i.e. this is a strategy that almost-surely wins the parity objective. Clearly, if $x > y$ then it is better to play b for optimizing the mean-payoff, otherwise, it is better to play a. As x and y are unknown, we need to learn estimates $\hat{x}$ and $\hat{y}$ for those values to make a decision. This can be done by playing a and b a number of times from q_0 and by observing how many times we get up and how many times we get down. If $\hat{x} > \hat{y}$, we may choose to play b forever in order to optimize our mean payoff. We then face two difficulties. First, after the learning episode, we may instead observe $\hat{x} < \hat{y}$ while $x \geq y$. This is because we may have been unlucky and observed statistics that differ from the real distribution. Second, playing b always is not an option if

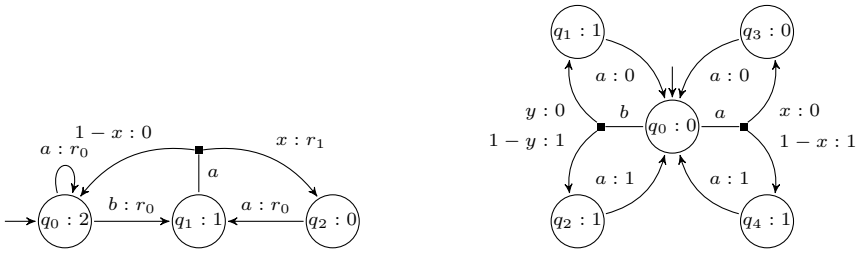

Fig. 1. Two automata, representing unknown MDPs, are depicted in the figure. Actions label edges from states (circles) to distributions (squares); a probability-reward pair, edges from distributions to states; an action-reward pair, Dirac transitions; a name-priority pair, states.

we want to satisfy the parity objective with probability 1 (almost surely). We give algorithms to overcome these two problems and compute a finite-memory strategy that satisfies the parity objective with probability 1 and is close to the optimal expected mean-payoff value with high probability.

The finite-memory learning strategy produced by our algorithm works as follows in this example. First, it chooses $n \in \mathbb{N}$ large enough so that trying a and b from q_0 as many as n times allows it to learn $\hat{x}$ and $\hat{y}$ such that $|\hat{x} - x| \leq \varepsilon$ and $|\hat{y} - y| \leq \varepsilon$ with probability at least $1 - \gamma$. Then, if $\hat{x} > \hat{y}$ the strategy plays b for K steps and then a once. K is chosen large enough so that the mean payoff of any run will be ε-close to the best obtainable expected mean payoff with probability at least $1 - \gamma$. Furthermore, as a is played infinitely many times, the upper-right state will be visited infinitely many times with probability 1. Hence, the strategy is also almost-surely satisfying the parity objective.

We also show that if we allow for learning all along the execution of the strategy then we can get, on this example, the exact optimal value and satisfy the parity objective almost surely. However, to do so, we need infinite memory.

For details, see the full version [4].

References

1. Krzysztof R. Apt and Erich Grädel. *Lectures in game theory for computer scientists.* Cambridge University Press, 2011.
2. T. Brázdil, K. Chatterjee, M. Chmelík, V. Forejt, J. Křetínský, M. Kwiatkowska, D. Parker, and M. Ujma. Verification of markov decision processes using learning algorithms. In *ATVA*, 2014.
3. P. Daca, T. Henzinger, J. Křetínský, and T. Petrov. Faster statistical model checking for unbounded temporal properties. In *TACAS*, 2016.
4. J. Křetínský, G.A. Pérez, and J.-F. Raskin. Learning-based mean-payoff optimization in an unknown MDP under omega-regular constraints. In *CONCUR*, 2018.
5. R.S. Sutton and A.G. Barto. *Reinforcement learning: an introduction.* Adaptive computation and machine learning. MIT Press, 2018.