An Anytime Algorithm for Reachability on Uncountable MDP

Kush Grover

Technische Universität München

The standard formalism for modelling systems with both non-deterministic and probabilistic behaviour are Markov decision processes (MDP) [1]. In the context of many applications such as cyber-physical systems, states and actions are used to model real-valued phenomena such as position or throttle. Consequently, the state space and the action space may be uncountably infinite. For example, a (multi-dimensional) real interval $[a, b] \times [c, d] \subseteq \mathbb{R}^2$ can model a safe area for a robot to move in or a set of available control inputs such as acceleration and steering angle. This gives rise to MDP with potentially uncountable state- and action-spaces (sometimes called controlled discrete-time Markov Process or discrete-time Markov Control Process), with applications ranging from modelling a Mars rover [2, 3], over water reservoir control, warehouse storage management, energy control and many more.

Although systems modelled by MDP are often safety-critical, the analysis of uncountable systems is so complex that practical approaches for verification and controller synthesis are based on unreliable 'best effort' learning techniques, for example *reinforcement learning*. While efficient in practice, these methods guarantee, even in the best case, convergence to the true result only in the limit, e.g., [4], or for increasingly precise discretization, e.g., [5]. In line with the tradition of learning and to make the analysis more feasible, the typical objectives considered for MDP are either finite-horizon [6, 7] or discounted properties [8], together with restrictive assumptions. Note that when it comes to approximation, discounted properties effectively are finite-horizon. In contrast, ensuring safety of a reactive system or a certain probability to satisfy its mission goals requires an *unbounded* horizon and reduces to optimizing the reachability probabilities. Moreover, the safety-critical context requires to give *reliable* bounds on the probability, not an approximation with *unknown* precision.

This work provides an algorithm for reachability on Markov decision processes with uncountable state and action spaces, which, under mild assumptions, approximates the optimal value to any desired precision. It is the first such *anytime* algorithm, meaning that at any point in time it can return the current approximation with its precision. Moreover, it is the first algorithm able to utilize *learning* approaches simultaneously without sacrificing *guarantees* and it further allows for combination with existing heuristics.

We first give an algorithm which extends *value iteration* (VI) [9, 1] to this uncountable setting. It converges to the true values under very mild assumptions. The second and the main algorithm combines this VI in general setting with *bounded real-time dynamic programming* (BRTDP). It is the first algorithm for reachability in such MDP with correct, converging bounds on the precision/error of the result, which furthermore is able to omit unimportant parts of the state space. The main idea of the algorithm is very similar to BRTDP: It samples some state-action pairs and keeps track of lower and upper bounds of those pairs, now it can estimate safe lower and upper bounds of the other pairs using Lipschitz continuity. The lower and upper bounds for all the state-action pair comes closer and closer as the algorithm samples more pairs. When the difference between lower and upper bound for the initial state becomes less than a given precision, the algorithm terminates. Since we have lower and upper bounds for all the state-action pairs at any given point, the algorithm can be queried at any point to get those bounds for the initial state making this an *anytime* algorithm. This algorithm can also use some existing learning heuristics to speed up the computation without sacrificing the guarantees. Additionally, we identify a rich, natural subclass of LTL to which our algorithms can be directly extended.

This is a joint work with Jan Křetínský, Tobias Meggendorfer and Maximilian Weininger.

References

[1] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Statistics. Wiley, 1994.

[2] John L. Bresina, Richard Dearden, Nicolas Meuleau, Sailesh Ramakrishnan, David E. Smith, and Richard Washington. Planning under continuous time and resource uncertainty: A challenge for AI. *CoRR*, abs/1301.0559, 2013.

[3] Mohammadhosein Hasanbeig, Alessandro Abate, and Daniel Kroening. Certified reinforcement learning with logic guidance. *CoRR*, abs/1902.00778, 2019.

[4] Francisco S. Melo, Sean P. Meyn, and M. Isabel Ribeiro. An analysis of reinforcement learning with function approximation. In William W. Cohen, Andrew McCallum, and Sam T. Roweis, editors, *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, volume 307 of *ACM International Conference Proceeding Series*, pages 664–671. ACM, 2008.

[5] Manfred Jaeger, Peter Gjøl Jensen, Kim Guldstrand Larsen, Axel Legay, Sean Sedwards, and Jakob Haahr Taankvist. Teaching stratego to play ball: Optimal synthesis for continuous space mdps. In Yu-Fang Chen, Chih-Hong Cheng, and Javier Esparza, editors, *Automated Technology for Verification and Analysis - 17th International Symposium, ATVA 2019, Taipei, Taiwan, October 28-31, 2019, Proceedings*, volume 11781 of *Lecture Notes in Computer Science*, pages 81–97. Springer, 2019.

[6] Lihong Li and Michael L. Littman. Lazy approximation for solving continuous finite-horizon mdps. In Manuela M. Veloso and Subbarao Kambhampati, editors, *Proceedings, The Twentieth National Conference on Artificial Intelligence and the Seventeenth Innovative Applications of Artificial Intelligence Conference, July 9-13, 2005, Pittsburgh, Pennsylvania, USA*, pages 1175–1180. AAAI Press / The MIT Press, 2005.

[7] Alessandro Abate, Maria Prandini, John Lygeros, and Shankar Sastry. Probabilistic reachability and safety for controlled discrete time stochastic hybrid systems. *Automatica*, 44(11):2724–2734, 2008.

[8] Carlos Guestrin, Milos Hauskrecht, and Branislav Kveton. Solving factored mdps with continuous and discrete variables. In David Maxwell Chickering and Joseph Y. Halpern, editors, *UAI '04, Proceedings of the 20th Conference in Uncertainty in Artificial Intelligence, Banff, Canada, July 7-11, 2004*, pages 235–242. AUAI Press, 2004.

[9] Ronald A Howard. Dynamic programming and markov processes. 1960.