Analysis of Learning and Planning with Options

Ronan Fruit, Alessandro Lazaric

INRIA Lille, 40, avenue Halley, 59650 Villeneuve d'Ascq
ronan.fruit@inria.fr and alessandro.lazaric@inria.fr

Markov Decision Processes (MDPs) are an expressive mathematical tool to model agents aiming at maximizing a long-term reward in a stochastic environment. Many efficient algorithms have been developed to solve both probabilistic planning problems (when the MDP is known) and online learning problems (when the MDP is unknown and need to be explored). In real-world applications however (like in robotics), algorithms such as Value Iteration (planning), R-MAX, E3, and UCRL (learning) suffer from the curse of dimensionality. Macro-actions (often called "options" (Sutton *et al.*, 1999)) and hierarchical structures (e.g., MAX-Q Decomposition (Dietterich, 2000)) were introduced to overcome this problem by providing additional knowledge to learning and planning algorithms and improve their performance. In planning, the improvement corresponds to a reduced computational complexity, whereas for learning, options may reduce the sample complexity or the regret. Despite their potential benefit, it has also been observed that options may actually be harmful (Jong *et al.*, 2008). This contrasting evidence on the impact of options on the planning/learning performance calls for a more thorough theoretical analysis of their properties.

Research questions. MDPs with options are usually formalized as Semi-Markov Decision Processes (SMDPs). SMDPs are MDPs with transitions taking a random amount of time (Puterman, 1994). While many algorithms used for MDPs can be transposed to SMDPs (e.g., dynamic programming and model-based learning algorithms), they often require additional information that is not directly available, such as the average duration of options, their average reward, and their transition probabilities. In planning, computing these quantities may come with an extra computational cost, while in learning, their estimation may require additional samples. As a result, the main question that arises is: given a fixed set of options, what is its overall impact on the planning/learning performance in a given MDP? Addressing this question would allow to understand in which MDPs a given set of options may lead to an actual performance improvement. A natural next step would be then to optimize the construction of options so as to maximize the improvement. While there is a wide body of literature on automated option discovery (Mcgovern & Barto, 2001; Menache *et al.*, 2002), a rigorous theoretical understanding of these approaches is still missing.

Preliminary results.

Analysis of SMDP learning algorithms. Since an MDP with options can be seen as an SMDP, it is natural to start analysing the theoretical properties of options using the existing theory on SMDPs. Brunskill & Li (2014) extend R-MAX and E3 to SMDPs in the total discounted reward setting and analyze their performance in the PAC-MDP framework. Inspired by this work, we designed an UCRL-based algorithm that can be used to learn an SMDP in the average reward setting and derived a regret bound for it. The bound naturally extends the MDP analysis of Jaksch *et al.* (2010). More precisely, the bound is of order $O(S(D + C_r + C_\tau + \bar{\tau}_{\max})\sqrt{An})$, where S and A are the number of states and actions, n is the number of decision steps, D extends the notion of diameter to the case of SMDPs, C_r and C_τ are constants depending on the distribution of rewards and duration of the actions, and $\bar{\tau}_{\max}$ is the maximal expected duration.

Analysis of the impact of options in MDPs. While the previous result gives a better understanding of the performance of learning in SMDPs, it provides very limited information of when options can be beneficial. A possible explanation is that considering an MDP with options as an SMDP might not be an appropriate approach to study the theoretical properties of options. When moving to SMDPs, the impact of options on the MDP dynamics is somehow lost. In fact, a running option is a Markov Reward Process with a stopping criterion whereas an action taken in an SMDP can be any arbitrary stopped random process. We thus introduced a new way of analysing MDPs with options that preserves the structure. The idea consists in building an extended MDP in which we still enforce the constraints of the options, while making sure that classical algorithms can still be applied. Exploiting this novel structure, we are able to derive explicit conditions on the structure of the set of options to ensure that the computational complexity per iteration of Value Iteration is decreased (at the expense of an increase in memory complexity).

References

BRUNSKILL E. & LI L. (2014). Pac-inspired option discovery in lifelong reinforcement learning. In T. JEBARA & E. P. XING, Eds., *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, p. 316–324: JMLR Workshop and Conference Proceedings.

DIETTERICH T. G. (2000). Hierarchical reinforcement learning with the maxq value function decomposition. *Journal of Artificial Intelligence Research*, **13**, 227–303.

JAKSCH T., ORTNER R. & AUER P. (2010). Near-optimal regret bounds for reinforcement learning. *J. Mach. Learn. Res.*, **11**, 1563–1600.

JONG N. K., HESTER T. & STONE P. (2008). The utility of temporal abstraction in reinforcement learning. In *The Seventh International Joint Conference on Autonomous Agents and Multiagent Systems*.

MCGOVERN A. & BARTO A. G. (2001). Automatic discovery of subgoals in reinforcement learning using diverse density. In *In Proceedings of the eighteenth international conference on machine learning*, p. 361–368: Morgan Kaufmann.

MENACHE I., MANNOR S. & SHIMKIN N. (2002). Q-cut - dynamic discovery of sub-goals in reinforcement learning. In *Machine Learning: ECML 2002, 13th European Conference on Machine Learning, Helsinki, Finland, August 19-23, 2002, Proceedings*, p. 295–306.

PUTERMAN M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York, NY, USA: John Wiley & Sons, Inc., 1st edition.

SUTTON R., PRECUP D. & SINGH S. (1999). Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, **112**, 181–211.