Rationality in General Reinforcement Learning

Peter Sunehag, Marcus Hutter
peter.sunehag@anu.edu.au, marcus.hutter@anu.edu.au
Research School of Computer Science, Australian National University
Canberra, Australia

May 9, 2013

The general reinforcement learning problem in unknown environments is an extremely challenging problem [Hut05]. Proving meaningful guarantees for an agent that directly says something positive about how the agent will perform and that holds with high probability or in expectation with respect to the true environment seems very difficult. We here suggest a framework where we combine notions of what is desirable in decision theory with some concepts from reinforcement learning. In this framework, an agent consists of a decision function and a hypothesis generating function. The hypothesis generating function feeds the decision function a class of environments at every time step and the decision function chooses an action/policy given such a class. The decision theoretic analysis is used to restrict the choice of the decision function. We consider asymptotic properties and error bounds when designing the hypothesis generating function.

In the foundations of decision theory, the focus is on axioms for rational preferences [NM44, Sav54]. The setting is usually that of a single decision and the decision is assumed to not affect the outcome of the observed event, but only its utility. The single decision setting can actually be understood as incorporating sequential decision making since the one choice can be for a policy to follow. This latter perspective is called normal formal in game theory. [SH11] extends rationality theory to the full reinforcement learning setting. It follows from the strictest version of these axioms that the agent must be a Bayesian agent [SH11]. In [SH12b], we slightly loosened the axioms in a way that is closely related to the multiple-prior setting from [GS89], except that it allows optimism instead of pessimism and is based on a given utility function. The optimism allows for better asymptotic optimality guarantees which is of interest in reinforcement learning.

In the field of reinforcement learning, there has been much work dedicated to designing agents for which one can prove asymptotic optimality or sample complexity bounds. The latter are high probability bounds on the number of time steps where the agent does not take a near optimal decision [SLL09]. Most of the work has been dedicated to the setting of discounted Markov Decision

Processes (MDPs) but recently also the case of classes of general environments has been addressed [SH12a, LHS13]. However, a weakness with sample complexity bounds is that they do not directly guarantee good performance for the agent since an agent who has the opportunity to self-destruct can achieve perfect prediction of its future and remove all uncertainty by choosing this option. Hence, aiming for the best sample complexity can be a very bad idea in general reinforcement learning. If one restricts oneself to settings like ergodic MDPs or value-stable environments [RH08] where the agent can always still achieve as high rewards as one could from the start, these bounds are directly meaningful. In the general case they are not. Here we consider agents that are rational in a decision theoretic sense and within that class design agents that make few errors.

References

[GS89] I. Gilboa and D. Schmeidler. Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18(2):141–153, April 1989.

[Hut05] M. Hutter. *Universal Articial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin, 2005.

[LHS13] T. Lattimore, M. Hutter, and P. Sunehag. The sample-complexity of general reinforcement learning. In *International Conference of Machine Learning (ICML'2013)*, 2013.

[NM44] J. Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

[RH08] D. Ryabko and M. Hutter. On the possibility of learning in reactive environments with arbitrary dependence. *Theor. Comput. Sci.*, 405(3):274–284, 2008.

[Sav54] L. Savage. *The Foundations of Statistics*. Wiley, New York, 1954.

[SH11] P. Sunehag and M. Hutter. Axioms for rational reinforcement learning. In *Algorithmic Learning Theory - 22nd International Conference, ALT 2011, Espoo, Finland, October 5-7, 2011. Proceedings*, volume 6925 of *Lecture Notes in Computer Science*, pages 338–352. Springer, 2011.

[SH12a] P. Sunehag and M. Hutter. Optimistic agents are asymptotically optimal. In *Proceedings of the 25:th Australasian AI conference*, 2012.

[SH12b] P. Sunehag and M. Hutter. Optimistic AIXI. In *Proceedings of the 4:th conference on Artifical General Intelligence (AGI'2012)*, pages 312–321, 2012.

[SLL09] A. L. Strehl, L. Li, and M. L. Littman. Reinforcement learning in finite MDPs: PAC analysis. *Journal of Machine Learing Research*, 10:2413–2444, 2009.