Reinforcement Learning: A Graduate Course (6hp)

Alexandre Proutiere

June 12, 2017

Abstract. Reinforcement Learning (RL) addresses the problem of controlling a dynamical system so as to maximize a notion of reward cumulated over time. At each time (or round), the agent selects an action, and as a result, the system state evolves. The agent observes the new state and collects a reward associated with the state transition, before deciding on the next action. Unlike classical control tasks where typically the system dynamics are completely predictable, RL is concerned with systems whose dynamics have to be learnt or with systems interacting with an uncertain environment. As time evolves, the agent gathers more data, and may improve her knowledge about the system dynamics to make better informed decisions. RL has found numerous applications, ranging from robotics, control, online services and game playing, and has received an increasing attention. Very recently, RL has solved problems in situations approaching real-world complexity, e.g., in learning human-level control for playing video and board games. These situations are however rather specific, and we are still far from systems able to learn in a wide variety of scenarios like humans do.

The course will provide an in-depth treatment of the modern theoretical tools used to devise and analyze RL algorithms. It includes an introduction to RL and to its classical algorithms such as Q-learning, and SARSA, but further presents the rationale behind the design of more recent algorithms, such as those striking optimal trade-off between exploration and exploitation (e.g. UCRL). The course also covers algorithms used in recent RL success stories, i.e., deep RL algorithms. Some basic notions in probability theory are required to follow the course.

1 Course Summary

L1. Introduction to Reinforcement Learning.

L2. Markov Decision Processes and Bellman's equation for finite and infinite horizon (with or without discount).

L3. RL problems. Regret, sample complexity, exploration-exploitation trade-off.

L4. First RL algorithms (e.g. Q-learning, TD-learning, SARSA). Convergence analysis.

L5. Bandit optimization: the "optimism in face of uncertainty" principle vs. posterior sampling

L6. RL algorithms 2.0 (e.g. UCRL, Thompson Sampling, REGAL). Regret and sample complexity analysis.

L7. Scalable RL algorithms: State aggregation, function approximation (deep RL, experience replay).

L8. Examples and empirical comparison of various algorithms.

2 Course Organisation

The course is organised as an intensive course divided into two blocks. This gives the participants time to digest the first material, while still providing opportunities to ask detailed questions. Specifically

1. General: (Tue.-Friday) 3,4,5,6 Oct, 10.15-12.00am.
2. Advanced: (Tue.-Friday) 17,18,19,20 Oct. 10.15-12.00am.

3 About the Lecturer

The lecturer is Alexandre Proutiere (`http://people.kth.se/~alepro/`), professor in Automatic Control at KTH, Royal Institute of Technology. He has being intensely working in online optimization and RL over the last few years, and has published his contributions in the three main conferences on the theory of machine learning (NIPS, ICML, and COLT). He is the local chair of COLT 2018, to be held in Stockholm.

Before joining KTH, Alexandre was a permanent researcher at Microsoft Research (Cambridge) from 2007 to 2011, a research engineer at France Telecom R&D from 2000 to 2006. He received his PhD in Applied Mathematics from Ecole Polytechnique, graduated in mathematiques from Ecole Normale Superieure, and has an engineering degree from Telecom Paris. He won the ACM Sigmetrics rising star award in 2009, and received the ACM best papers awards at Sigmetrics 2004 and 2010, and Mobihoc 2009. He holds an ERC consolidator grant on bandit optimization and its applications.

4 Course Material

The course is not based on a single textbook, but builds on a series of key publications in the field. Here is a preliminary list of relevant articles and books.

[1] P. Auer, T. Jaksch, and R. Ortner. Near-optimal regret bounds for reinforcement learning. In Advances in Neural Information Processing Systems 22, pages 8996, 2009.

[2] A. Burnetas and M. Katehakis. Optimal adaptive policies for sequential allocation problems. Advances in Applied Mathematics, 17(2):122142, 1996.

[3] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. Human-level control through deep reinforcement learning. Nature, 2015.

[4] M. Puterman. Markov Decision Processes: Discrete Stochastic Dynamic Programming. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1994.

[5] R. Sutton and A. Barto. Introduction to Reinforcement Learning. MIT Press, Cambridge, MA, USA, 1st edition, 1998

[6] C. Szepesvri. Algorithms for Reinforcement Learning. Synthesis Lectures on Articial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2010.

[7] J. Tsitsiklis and B. Roy. An analysis of temporal-difference learning with function approximation. IEEE Transactions on Automatic Control, 1997.

[8] C. Watkins. Learning from delayed rewards. PhD thesis, University of Cambridge England, 1989.

[9] M. Wiering and M. van Otterlo. Reinforcement Learning: State-of-the-Art. Springer, 2014.

Course material (slides, publication, news, ...) will be made available via the website of the course:

`http://www.it.uu.se/research/systems_and_control`

Contact Kristiaan Pelckmans (kp@it.uu.se) for further questions.