Carnegie Mellon University

THE ROBOTICS INSTITUTE

THESIS PROPOSAL

Towards Generalization and Efficiency of Reinforcement Learning

Thursday, March 8, 2018

4405 Gates Hillman Center

10:00 a.m.

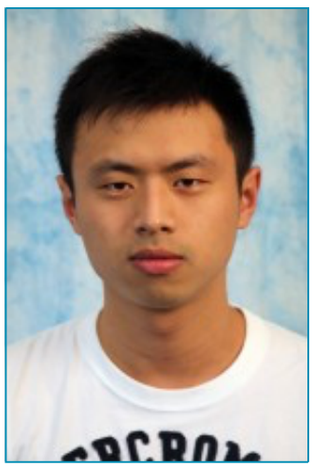
Thesis Committee:

J. Andrew Bagnell
Chair

Geoff J. Gordon

Martial Hebert

Byron Boots
Georgia Institute of Technology



Wen Sun

Abstract

In classic supervised machine learning, a learning agent behaves as a passive observer: it receives examples from some external environment which it has no control over and then makes predictions. The predictions the agent made will not affect any future examples it will see (i.e., examples are identically and independently sampled from some unknown distribution). That is, the agent will not actively intervene the environment. Reinforcement Learning, on the other hand, is fundamentally *interactive* : an autonomous agent must learn how to behave in an unknown, uncertain, and possibly hostile environment, by actively interacting with the environment to collect useful feedback to improve its sequential decision making ability. The RL agent will also intervene in the environment: the agent makes decisions which in turn affects further evolution of the environment.

Because of its generality– most machine learning problems can be viewed as special cases– RL is hard. As there is no direct supervision, one central challenge in RL is how to explore an unknown environment and collect useful feedback *efficiently*. In recent RL success stories (e.g., super-human performance on video games [Mnih et al., 2015]), we notice that most of them rely on *random exploration strategies*, such as ε-greedy. While ε-greedy is simple and asymptotically optimal, it requires large number of interactions with the environment before it can learn anything useful. Similarly, policy gradient method such as REINFORCE [Williams, 1992], perform exploration by injecting randomness into action space and hope the randomness can lead to a good sequence of actions that achieves high total reward. The theoretical RL literature has developed more sophisticated algorithms for efficient exploration (e.g., [Azar et al., 2017]), however, the sample complexity of these near-optimal algorithms *has* to scale exponentially with respect to key parameters of underlying systems such as dimensions of state and action space. Such exponential dependence prohibits a direct application of these theoretically elegant RL algorithms to large-scale applications. In summary, without any further assumptions, RL is hard, both in practice and in theory.

In this work, we attempt to gain purchase on the RL problem by introducing additional assumptions and sources of information, and then *reducing the resulting RL problem to simpler problems* which we understand well and know how to solve. The first contribution of this work comes from a *reduction of policy evaluation to no-regret online learning*. As no-regret online learning is an active research area that has well-established theoretical foundation and appealing practical usage, such a reduction creates a new family of algorithms for provably correct policy evaluation under **very weak** assumptions on the generating process. This enables us to have correct boostrapping policy evaluation algorithms without the requirement of a Markov process. Further the reduction allows any new, faster no-regret online algorithm to immediately translate to a faster policy evaluation algorithm. The second contribution of this work comes from *improving RL sample efficiency via Imitation Learning (IL)*. Imitation Learning reduces policy improvement to classic supervised learning, which is a well-established research area and has provably correct algorithms along with efficient implementations. We study in both theory and in practice how one can imitate experts to reduce sample complexity compared to a pure RL approach. The third contribution of this work comes from leveraging efficient model-base optimal control by using a *reduction of RL to IL*. We explore the possibilities of learning local models and then using model-based optimal control solvers to compute an intermediate "expert" for efficient policy improvement via imitation. We propose a general framework, named *Dual Policy Iteration (DPI)*, which maintains two policies, an apprentice policy and an expert policy, and alternatively updates the apprentice policy via imitating the expert policy while updates the expert policy via model-based optimal control with a learned local model. Furthermore, we show a general convergence analysis that extends the existing approximate policy iteration theories to DPI. DPI generalizes and provides the first theoretical foundation for recent successful practical RL algorithms such as ExIt and AlphaGo Zero [Anthony et al., 2017, Silver et al., 2017], and provides a theoretical sound and practically efficient way of unifying model-based and model-free RL approaches.