Stable Reinforcement Learning with Unbounded State Space (Extended Abstract)

Devavrat Shah DEVAVRAT@MIT.EDU
LIDS, MIT

Qiaomin Xie QIAOMIN.XIE@CORNELL.EDU
ORIE, Cornell University

Zhi Xu ZHIXU@MIT.EDU
LIDS, MIT

Editors: A. Bayen, A. Jadbabaie, G. J. Pappas, P. Parrilo, B. Recht, C. Tomlin, M.Zeilinger

We consider the problem of reinforcement learning (RL) for controlling an unknown dynamical systems with an unbounded state space. Such problems are ubiquitous in various application domains, as exemplified by scheduling for queueing networks. As a paradigm for learning to control dynamical systems, RL has a rich literature. In particular, algorithms for settings with finite/compact state spaces has been well studied, with both classical asymptotic convergence results and recent non-asymptotic performance guarantees. However, literature on problems with unbounded state space is scarce, with exception on settings with special and *known* structures, such as linear quadratic regulator problems.

The unboundedness of the state space poses new challenges in both algorithm design and analysis. First, we argue that an RL approach that relies on *offline* training only is bound to fail. As the state space is unbounded, with a non-zero probability the system will reach a state that is not previously observed in the *finite* training data. The deployed policy is likely to have undesirable behaviors for such unseen states. Therefore, to learn a reasonable policy with an unbounded state space, we need to consider *online* policies that update whenever a new scenario is encountered. Second, we need a new measure to quantify the desired performance of such online policies. Because of the unbounded state space, expecting a good approximation of the optimal value function over the entire state space is not a meaningful measure — an alternative notion that quantifies the "goodness" of policies under such setting is desirable.

Motivated by the above considerations, we study discounted Markov Decision Processes with an unbounded state space and a finite action space. As our main contribution, we propose *stability* as the notion of "goodness" of RL policy, inspired by literature in queueing systems and control theory. Informally, an RL policy is stable if the resulting state dynamics remain in a *bounded* region with high probability. As a proof of concept, we present an RL policy using Sparse-Sampling-based Monte Carlo oracle and show that it is stable, as long as the system dynamics under the optimal policy respects a Lyapunov function with drift condition. The assumption of existence of a Lyapunov function is not restrictive as it is equivalent to the positive recurrence or stability property of any Markov chain, i.e., if there is any policy that can stabilize the system, then it must posses a Lyapunov function. Also, our policy *does not* require knowledge of Lyapunov function. It is our analysis that uses the Lyapunov function and establishes that the probability of the state being away from a bounded region decays exponentially with the "distance" to the bounded region.