Scalable Reinforcement Learning for Multi-Agent Networked Systems

GUANNAN QU and ADAM WIERMAN, California Institute of Technology, USA
NA LI, Harvard University, USA

1 EXTENDED ABSTRACT

The modeling and optimization of networked systems such as wireless communication networks and traffic networks is a long-standing challenge. Typically analytic models must make numerous assumptions to obtain tractable models as a result of the complexity of the systems, which include many unknown, or unmodeled dynamics. Given the success of Reinforcement Learning (RL) in a wide array of domains, it has emerged as a promising tool for tackling the complexity of networked systems. However, when seeking to use RL in the context of the control and optimization of large-scale networked systems, scalability quickly becomes an issue. The goal of this paper is to develop *scalable* multi-agent RL for networked systems.

Motivated by real-world networked systems like wireless communication, epidemics and traffic, we considier a RL model of n agents with *local interaction structure*. Specifically, each agent i has local state s_i, local action a_i and the agents are associated with an underlying dependence graph $\mathcal{G}$ and interact locally, i.e, the distribution of $s_i(t+1)$ only depends on the current states of the local neighborhood of i as well as the local $a_i(t)$. Further, each agent is associated with stage reward r_i that is a function of s_i, a_i, and the global stage reward is the average of r_i. In this setting, the design goal is to find a decision policy that maximizes the (discounted) global reward. This setting captures a wide range of applications, e.g. epidemics [9], social networks [4], wireless communication networks [13].

A fundamental difficulty when applying RL to such networked systems is that, even if individual state and action spaces are small, the entire state profile $(s_1, \ldots, s_n)$ and the action profile $(a_1, \ldots, a_n)$ can take values from a set of size exponentially large in n. This "curse of dimensionality" renders the problem unscalable. For example, most RL algorithms such as temporal difference (TD) learning or Q-learning require storage of a Q-function [1] whose size is the same as the state-action space, which is exponentially large in n. Such scalability issues have indeed been observed in previous research on variants of the problem we study, e.g. in multi-agent RL [3, 8] and factored Markov Decision Proccess (MDP) [6, 7]. A variety of approaches have been proposed to manage this issue, e.g. the idea of "independent learners" in [5, 11]; or function approximation schemes [12]. However, such approaches lack rigorous optimality guarantees. In fact, it has been suggested that such MDPs with exponentially large state spaces may be fundamentally intractable, e.g., see [2].

In addition to the scalability issue, another challenge is that, even if an optimal policy that maps a global state $(s_1, \ldots, s_n)$ profile to a global action $(a_1, \ldots, a_n)$ can be found, it is usually impractical to implement such a policy for real-world networked systems because of the limited information and communication among agents. For example, in large scale networks, each agent i may only be able to to implement *localized policies*, where its action a_i only depends on its own state s_i. Designing such localized polices with global network performance guarantee can also be challenging [10].

The challenges described above highlight the difficulty of applying RL to control large scale networked systems; however, the network itself provides some structure, particularly the local interaction structure, that can potentially be exploited. The question that motivates this paper

Authors' addresses: Guannan Qu, gqu@caltech.edu.com; Adam Wierman, adamw@caltech.edu, California Institute of Technology, USA; Na Li, nali@seas.harvard.edu, Harvard University, USA.

is: *Can the network structure be utilized to develop scalable RL algorithms that provably find a (near-)optimal localized policy?*

Contributions. In this work we propose a framework that exploits properties of the network structure to develop RL to learn localized policies for large-scale networked systems in a scalable manner. Specifically, our main result shows that our algorithm, Scalable Actor Critic (SAC), finds a localized policy that is a $O(\rho^{\kappa+1})$-approximation of a stationary point of the objective function, with complexity that scales with the local state-action space size of the largest κ-hop neighborhood. To the best of our knowledge, our results are the first to provide such provable guarantee for scalable RL of localized policies in multi-agent network settings.

The key technique underlying our results is we prove that, under the local interaction structure, the Q-function satisfies an *exponential decay property*, where the Q-function's dependence on far away nodes shrink exponentially in their graph distance with rate $\rho \leq \gamma$, where γ is the discounting factor. This leads to a tractable approximation of the Q-function. In particular, despite the Q-function itself being intractable to compute due to the large state-action space size, we introduce a *truncated Q-function* which only depends on a small spatial horizon, that can be computed efficiently and can be used in an actor-critic framework which yields an $O(\rho^{\kappa})$-approximation. This technique is novel and is a contribution in its own right. It can be used broadly to develop RL for network settings beyond the specific actor-critic algorithm we propose in this paper.

REFERENCES

[1] Dimitri P Bertsekas and John N Tsitsiklis. 1996. *Neuro-dynamic programming*. Vol. 5. Athena Scientific Belmont, MA.

[2] Vincent D Blondel and John N Tsitsiklis. 2000. A survey of computational complexity results in systems and control. *Automatica* 36, 9 (2000), 1249–1274.

[3] Lucian Bu, Robert Babu, Bart De Schutter, et al. 2008. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 38, 2 (2008), 156–172.

[4] Deepayan Chakrabarti, Yang Wang, Chenxi Wang, Jurij Leskovec, and Christos Faloutsos. 2008. Epidemic thresholds in real networks. *ACM Transactions on Information and System Security (TISSEC)* 10, 4 (2008), 1.

[5] Caroline Claus and Craig Boutilier. 1998. The dynamics of reinforcement learning in cooperative multiagent systems. *AAAI/IAAI* 1998 (1998), 746–752.

[6] Carlos Guestrin, Daphne Koller, Ronald Parr, and Shobha Venkataraman. 2003. Efficient solution algorithms for factored MDPs. *Journal of Artificial Intelligence Research* 19 (2003), 399–468.

[7] Michael Kearns and Daphne Koller. 1999. Efficient reinforcement learning in factored MDPs. In *IJCAI*, Vol. 16. 740–747.

[8] Michael L Littman. 1994. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*. Elsevier, 157–163.

[9] Wenjun Mei, Shadi Mohagheghi, Sandro Zampieri, and Francesco Bullo. 2017. On the dynamics of deterministic epidemic propagation over networks. *Annual Reviews in Control* 44 (2017), 116–128.

[10] Michael Rotkowitz and Sanjay Lall. 2005. A characterization of convex problems in decentralized control. *IEEE transactions on Automatic Control* 50, 12 (2005), 1984–1996.

[11] Ming Tan. 1993. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the tenth international conference on machine learning*. 330–337.

[12] John N Tsitsiklis and Benjamin Van Roy. 1997. Analysis of temporal-diffference learning with function approximation. In *Advances in neural information processing systems*. 1075–1081.

[13] Alessandro Zocca. 2019. Temporal starvation in multi-channel CSMA networks: an analytical framework. *Queueing Systems* 91, 3-4 (2019), 241–263.