Global Convergence of Localized Policy Iteration in Networked Multi-Agent Reinforcement Learning

Yizhou Zhang*
Tsinghua University
China

Guannan Qu*
Carnegie Mellon University
United States

Pan Xu*
Duke University
United States

Yiheng Lin
California Institute of Technology
United States

Zaiwei Chen
California Institute of Technology
United States

Adam Wierman
California Institute of Technology
United States

ABSTRACT

We study a multi-agent reinforcement learning (MARL) problem where the agents interact over a given network. The goal of the agents is to cooperatively maximize the average of their entropy-regularized long-term rewards. To overcome the curse of dimensionality and to reduce communication, we propose a Localized Policy Iteration (LPI) algorithm that provably learns a near-globally-optimal policy using only local information. In particular, we show that, despite restricting each agent's attention to only its κ-hop neighborhood, the agents are able to learn a policy with an optimality gap that decays polynomially in κ. In addition, we show the finite-sample convergence of LPI to the global optimal policy, which explicitly captures the trade-off between optimality and computational complexity in choosing κ. Numerical simulations demonstrate the effectiveness of LPI. This extended abstract is an abridged version of [12].

ACM Reference Format:
Yizhou Zhang, Guannan Qu, Pan Xu, Yiheng Lin, Zaiwei Chen, and Adam Wierman. 2023. Global Convergence of Localized Policy Iteration in Networked Multi-Agent Reinforcement Learning. In *Abstract Proceedings of the 2023 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS '23 Abstracts), June 19–23, 2023, Orlando, FL, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3578338.3593545

1 INTRODUCTION

Reinforcement learning (RL) has seen remarkable successes in recent years, many of which fall into the multi-agent setting, such as playing multi-agent games [5, 8], smart grid [1], queueing networks [10], etc. In this work, we focus on a form of *networked* Multi-Agent RL (MARL) where the agents interact according to a given network graph.

Compared to single-agent RL, MARL faces many additional challenges. First of all, the curse of dimensionality (which is already a major challenge in single-agent RL) becomes a more severe issue in MARL because the complexity of the problem in term of the size of state and action space scales exponentially with the number of agents. Moreover, since the agents are coupled by the global state, the training process requires extensive communication among the agents in the entire network.

To overcome the aforementioned challenges, the existing literature considers performing MARL with only local information. For example, it has been proposed that each agent's policy depends only on the states of itself and, potentially, its neighboring agents. An example is the recent work of Lin et al. [4], Qu et al. [6, 7], which focuses on learning localized policies where each agent is allowed to choose its action based on its own and neighbors' states. Beyond the localization in policy, it has also been proposed to approximate each agent's value or Q functions in a way that they only depend on the local and nearby agents' states and actions [3, 4, 6, 7, 11], as opposed to the full state. These approaches greatly relieve the computation and communication burden both empirically and theoretically.

Despite notable progress, there still exist significant limitations. As discussed above, many previous works, e.g., [2, 4, 6, 7, 9], have primarily focused on localized policies where each agent makes decisions only based on the local state of the agent and potentially, its neighbors, as opposed to the global state. This leads to a fundamental performance gap between the class of localized policies considered and the optimal centralized policy. Even when the best localized policy can be identified, there is a performance degradation compared to the best centralized policy, as the centralized policy class is a strict superset of the localized policy class. An important open question that remains is the following:

How large is the gap between localized policies and the optimal centralized policy?
How much information must be available to each local agent in order to achieve a near-optimal performance?

In this work, we provide the first bounds on the gap between localized and centralized policies under a MARL model in networked systems.

Another limitation of prior work studying MARL is that, while existing results have provided convergence bounds for local policies, the convergence of their methods is typically only to a suboptimal policy in the localized policy class. This is unsatisfying as converging to a stationary point does not even guarantee converging to the best localized policy. Therefore, another important and open question that remains is the following:

Is it possible to design a MARL algorithm that provably finds a near-globally-optimal policy using only local information?

*Yizhou Zhang, Guannan Qu, Pan Xu contributed equally to this work.

SIGMETRICS '23 Abstracts, June 19–23, 2023, Orlando, FL, USA

ACM ISBN 979-8-4007-0074-3/23/06.
https://doi.org/10.1145/3578338.3593545

This question has received increasing attention in single-agent settings but it is still open in the context of learning localized policies in MARL and in this work, we provide an affirmative answer to this question under an MARL model in networked systems.

1.1 Contributions

Motivated by the open questions above, our work proposes and analyzes a new class of localized policies for networked MARL and proposes a Localized Policy Iteration (LPI) algorithm that converges to a near-globally-optimal policy. Our main contributions are summarized in the following.

- **Near-Globally-Optimal κ-Hop Localized Policies.** We show that a class of κ-hop localized policies are nearly globally optimal, where a κ-hop localized policy means that each agent is allowed to choose its action based on the states of its κ-hop local neighborhood. More specifically, we show that there exists a κ-hop localized policy whose optimality gap is polynomially small in κ, where the optimality gap is with respect to the best centralized policy that is allowed to depend on the global state. As a result, even with a small κ, the class of κ-hop localized policies is near optimal despite the fact that each agent only uses information from within a small κ-hop neighborhood to make its decision. This result justifies using localized policies in networked MARL.
- **Localized Policy Iteration.** Motivated by the result described above, we propose a localized MARL algorithm, a.k.a. LPI. At a high level, LPI iteratively performs policy improvement (and policy evaluation), but is restricted to κ-hop policies. While standard policy improvement requires using the information of global states and actions in order to conduct the policy improvement step, we develop a *soft* policy improvement that approximately performs policy improvement to κ-hop localized policies using only local information. Therefore, our proposed algorithm can be implemented in a truly localized manner.
- **Finite-Sample Analysis of LPI.** We provide a global convergence guarantee for LPI, which holds for any policy evaluation method used as a subroutine in the evaluation step of LPI (denoted as `PolicyEvaluation`) that satisfies a mild condition. Furthermore, we provide the finite-sample complexity for LPI when choosing a specific `PolicyEvaluation` method proposed by Lin et al. [4]. Specifically, we show that in order to achieve an ε optimality (compared to the best centralized policy), one needs to use a κ-hop localized policy with $\kappa = \Theta(poly(\frac{1}{\varepsilon}))$ in LPI, and the sample complexity in learning such a κ-hop localized policy with ε optimality gap scales polynomially with the largest state-action space size of local neighborhoods, as opposed to the global network.

The key technical novelty in this paper is a policy closure argument, where we identify a class of policies that satisfy a form of spatial decaying properties and show that they are closed under the entropy regularized Bellman operator in MARL. This key observation implies that when starting from a policy in this class with spatial decaying properties, the policy improvement procedure will result in a new policy within this class, which further reveals that the optimal policy is also in the spatial decaying policy class. We then show that LPI is an inexact version of the above policy iteration given by the regularized Bellman operator, where the learned policies and Q-functions are truncated to only depend on a localized neighborhood of each agent, and the exact policy evaluation and improvement steps are approximated by finite-sample estimations. Therefore, combining the policy closure argument for the exact version of LPI and error bound analyses for the approximation made in LPI, we are able to show its convergence to a near global optimal policy, even though we only use localized information in LPI.

While the main contribution of this work is to theoretically show that LPI can converge to the global optimal policy using only localized information, we also verify its empirical performance on a simulated networked MARL problem. In particular, we design an example of a spreading process over a network where the optimal policy depends on the global states of each agent (not just the local information). We run LPI on this example and the results highlight that LPI can perform well even outside the region suggested by the theory.

REFERENCES

[1] Xin Chen, Guannan Qu, Yujie Tang, Steven Low, and Na Li. 2022. Reinforcement learning for selective key applications in power systems: Recent advances and future challenges. *IEEE Transactions on Smart Grid* (2022).

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

[3] 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.

[4] Yiheng Lin, Guannan Qu, Longbo Huang, and Adam Wierman. 2021. Multi-agent reinforcement learning in stochastic networked systems. *Advances in Neural Information Processing Systems* 34 (2021), 7825–7837.

[5] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (2015), 529.

[6] Guannan Qu, Yiheng Lin, Adam Wierman, and Na Li. 2020. Scalable multi-agent reinforcement learning for networked systems with average reward. *Advances in Neural Information Processing Systems* 33 (2020), 2074–2086.

[7] Guannan Qu, Adam Wierman, and Na Li. 2020. Scalable reinforcement learning of localized policies for multi-agent networked systems. In *Learning for Dynamics and Control*. PMLR, 256–266.

[8] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *nature* 529, 7587 (2016), 484.

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

[10] Neil Walton and Kuang Xu. 2021. Learning and information in stochastic networks and queues. In *Tutorials in Operations Research: Emerging Optimization Methods and Modeling Techniques with Applications*. INFORMS, 161–198.

[11] Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. 2018. Mean field multi-agent reinforcement learning. In *International conference on machine learning*. PMLR, 5571–5580.

[12] Yizhou Zhang, Guannan Qu, Pan Xu, Yiheng Lin, Zaiwei Chen, and Adam Wierman. 2023. Global convergence of localized policy iteration in networked multi-agent reinforcement learning. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 7, 1 (2023), 1–51.