Distributed Reinforcement Learning for Decentralized Linear Quadratic Control: A Derivative-Free Policy Optimization Approach (Extended Abstract)*

Yingying Li† YINGYINGLI@G.HARVARD.EDU
Yujie Tang† YUJIETANG@SEAS.HARVARD.EDU
Runyu Zhang RUNYUZHANG@FAS.HARVARD.EDU
Na Li NALI@SEAS.HARVARD.EDU
John A. Paulson School of Engineering and Applied Sciences, Harvard University.

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

Motivated by recent advances on reinforcement learning for centralized linear quadratic regulators, this paper studies distributed reinforcement learning for decentralized linear quadratic control. Specifically, we consider a linear system operated by N agents, each of which only has access to partial state observations, local actions, local costs, and limited communication capacity via a network at each stage. The goal is to minimize the infinite-horizon averaged global cost.

We propose a Zero-Order Distributed Policy Optimization algorithm (ZODPO) that learns local controllers in a distributed fashion. ZODPO leverages the ideas of policy gradient, zero-order optimization and consensus algorithms. In each iteration of ZODPO, the agents first estimate the global cost by a consensus-based method; each agent then uses the locally estimated global cost to form a zero-order partial gradient estimator with respect to the local controller parameters; finally, each agent conducts local policy gradient updates using the estimated partial gradient in parallel.

Further, we investigate the nonasymptotic performance of ZODPO for linear static local controllers. We show that all output controllers $\mathsf{K}(1), \ldots, \mathsf{K}(T_G)$ of ZODPO are stabilizing with high probability when the algorithmic parameters are properly chosen. In addition, due to the nonconvexity of the decentralized control problem, we measure the optimality of the output controllers by the squared norms of their gradients, and show that in order to achieve

$$\frac{1}{T_G}\sum_{s=1}^{T_G} \|\nabla J(\mathsf{K}(s))\|^2 \leq \epsilon,$$

for a sufficiently small tolerance $\epsilon > 0$, ZODPO requires a sample complexity of

$$\Theta\Big(\frac{n_K^3}{\epsilon^4}\max\Big\{n\beta_0^2, \frac{N}{1-\rho_W}\Big\}\Big),$$

where n_K denotes the dimension of the controller parameter K, n denotes the dimension of the global state, β_0 is a constant determined by the system, ρ_W captures the rate of consensus via the communication network. Notice that the sample complexity has polynomial dependence on the dimensions of interest, demonstrating the scalability of ZODPO.

Lastly, we provide numerical results of ZODPO on multi-zone HVAC systems.

* This work was supported by NSF CAREER 1553407, AFOSR YIP, ONR YIP, and OSR-2019-CoE-NEOM-4178.10. The full version can be found at https://arxiv.org/abs/1912.09135.

† The first two authors contribute equally.