Proceedings of Machine Learning Research vol 144:1–2, 2021

Provably Sample Efficient Reinforcement Learning in Competitive Linear Quadratic Systems

Jingwei Zhang JZHANGEY@CSE.UST.HK
Hong Kong University of Science and Technology

Zhuoran Yang ZY6@PRINCETON.EDU
Princeton University

Zhengyuan Zhou ZZHOU@STERN.NYU.EDU
New York University

Zhaoran Wang ZHAORANWANG@GMAIL.COM
Northwestern University

Abstract

We study the infinite-horizon zero-sum linear quadratic (LQ) games, where the state transition is linear and the cost function is quadratic in states and actions of two players. In particular, we develop an adaptive algorithm that can properly trade off between exploration and exploitation of the unknown environment in LQ games based on the optimism-in-face-of-uncertainty (OFU) principle. We show that (i) the average regret of player 1 (the min player) can be bounded by $\widetilde{\mathcal{O}}(1/\sqrt{T})$ against any fixed linear policy of the adversary (player 2); (ii) the average cost of player 1 also converges to the value of the game at a sublinear $\widetilde{\mathcal{O}}(1/\sqrt{T})$ rate if the adversary plays adaptively against player 1 with the same algorithm, i.e., with self-play. To the best of our knowledge, this is the first time that a probably sample efficient reinforcement learning algorithm is proposed for zero-sum LQ games.

Keywords: Exploration; Linear-Quadratic Game; Reinforcement Learning; Optimal Control

Introduction. Despite the extensive literature on LQR control, most of them focus on the single-agent setting, where there is only one player who makes sequential decisions in a stochastic environment. In many real-world learning and control problems, the existence of a disturbance or adversary is often inevitable because of modeling errors and differences in training and test scenarios (Pinto et al. (2017)). Nevertheless, the adversarial/game-theoretic setting of LQ control has rarely been studied.

In this paper, we consider the infinite-horizon zero-sum linear quadratic (LQ) games with unknown model parameters. In such a game, there are two players which takes actions simultaneously and their actions, together with the state of the environment, determine the immediate cost and the next state. Specifically, the cost function is quadratic in the current state and joint actions and the state transition is governed by a linear dynamic system. Moreover, such a game is zero-sum and thus the sum of rewards received by the two players are equal to zero. The goal of each player is to maximize its average cost in the presence of the opponent. Thus, from the perspective of player 1, we can formulate the LQ game as minimax optimization, where the goal of player 1 (player 2) is to minimize (maximize) the average cost of player 1. LQ games is one the simplest extension of LQR to the multi-agent and competitive setting, and it is closely related to robust control and risk-sensitive control (Whittle, 1981; Basar and Olsder, 1999; Başar and Bernhard, 2008). It is shown that the Nash equilibrium of the LQ game is captured by generalized Algebraic Riccati equation and the Nash

equilibrium policies for both players are linear functions of the state. For solving such a game, in addition to directly solving the Riccati equation, (Zhang et al., 2019a,b; Bu et al., 2019) characterize the landscape of the policy optimization problem and show that policy gradient methods provably find the Nash equilibrium policies efficiently. However, these results are all population analyses and require the knowledge of the model parameters. When the model is unknown, it remains open how to design a sample efficient algorithm which provably achieve the Nash equilibrium of the LQ game.

To tackle such a challenge, we focus on the LQ game under the online setting, where the sample efficiency is characterized by the notion of regret. We propose an adaptive control algorithm for player 1 based on the OFU principle and analyze her (average) regret defined by the difference of the average cost up to time T and the (optimal) value of the game, which is the minimum cost achievable by player 1, given that the adversary (player 2) also tries to maximize its own reward against player 1. Moreover, we prove that the adaptive algorithm based on the OFU principle can properly trade-off between exploration and exploitation and achieves an $\widetilde{\mathcal{O}}(1/\sqrt{T})$ regret for player 1 against both non-adaptive (including the dynamics-omniscient adversary) and adaptive adversaries. To the best of our knowledge, we propose the first provably sample efficient reinforcement learning algorithm for online LQ games with unknown model parameters.

References

Tamer Başar and Pierre Bernhard. *H-infinity optimal control and related minimax design problems: a dynamic game approach*. Springer Science & Business Media, 2008.

Tamer Basar and Geert Jan Olsder. *Dynamic noncooperative game theory*, volume 23. Siam, 1999.

Jingjing Bu, Lillian J Ratliff, and Mehran Mesbahi. Global convergence of policy gradient for sequential zero-sum linear quadratic dynamic games. *arXiv preprint arXiv:1911.04672*, 2019.

Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2817–2826. JMLR. org, 2017.

Peter Whittle. Risk-sensitive linear/quadratic/gaussian control. *Advances in Applied Probability*, pages 764–777, 1981.

Kaiqing Zhang, Bin Hu, and Tamer Başar. Policy optimization for $\mathcal{H}_2$ linear control with $\mathcal{H}_\infty$ robustness guarantee: Implicit regularization and global convergence. *arXiv preprint arXiv:1910.09496*, 2019a.

Kaiqing Zhang, Zhuoran Yang, and Tamer Basar. Policy optimization provably converges to Nash equilibria in zero-sum linear quadratic games. In *Advances in Neural Information Processing Systems*, pages 11602–11614, 2019b.