Contribution ID: **49** Type: **not specified**

On the Global Convergence of Policy Based Methods in Average Reward Problems

Monday, June 17, 2024 2:30 PM (30 minutes)

In the context of average reward Markov Decision Processes (MDPs), traditional approaches for obtaining performance bounds based on discounted reward formulations fail to provide meaningful bounds due to their dependence on the horizon. This limitation arises because average reward problems can be viewed as discounted reward problems, with the discount factor approaching 1, effectively extending the horizon to infinity. Consequently, theoretical convergence guarantees in the discounted reward framework scale unfavorably with the horizon length, yielding unbounded performance estimates. Therefore, obtaining meaningful convergence bounds for widely employed algorithms in the context of average reward MDPs has been an open problem.

In this study, we progress on two classes of algorithms tailored to the average reward objective. First, we examine policy-based reinforcement learning (RL) algorithms, which can be viewed as instances of approximate policy iteration (API). We provide finite time performance bounds of API and show that the asymptotic error goes to zero in the limit as policy evaluation and policy improvement errors tend to zero. We further cast several RL algorithms in the API framework to obtain their overall performance bounds. Second, we study the global convergence analysis of policy gradient algorithms in tabular ergodic average reward MDPs. We obtain a sublinear rate of convergence of the iterates to the globally optimal policy. Unlike discounted reward problems, where the discount factor acts as a source of contraction aiding convergence analysis, average reward problems lack this property. To tackle these challenges, we employ new methods of analysis to prove the global convergence of both classes of algorithms. These findings shed light on the convergence behavior of policy-based RL algorithms and pave the way for their practical application in average reward scenarios.

Primary author: MURTHY, Yashaswini (University of Illinois Urbana Champaign)

Co-authors: Mr KUMAR, Navdeep (Technion); Mr SHUFARO, Itai (Technion); Prof. MOHARRAMI, Mehrdad (University of Iowa); Prof. LEVY, Kfir (Technion); Prof. SRIKANT, Rayadurgam (UIUC); Prof. MANNOR, Shie (Technion)

Presenter: MURTHY, Yashaswini (University of Illinois Urbana Champaign)

Session Classification: Parallel session: Policy gradient methods: optimization and convergence