A New Challenge in Policy Evaluation

Shangtong Zhang

University of Virginia
85 Engineer's Way, Charlottesville, VA, 22903, USA
shangtong@virginia.edu

Introduction

Reinforcement Learning (RL) builds agents that make decisions in multiple consecutive stages to achieve some goal. Enormous success of RL has been witnessed in simulated environments, e.g., Go and StarCraft II. Our ultimate goal is, however, to deploy RL agents in real-world scenarios, e.g., autonomous vehicles. Before such deployment, we have to know how good an RL agent is, ideally through some numeric metrics. This is the *policy evaluation* problem. Evaluating an agent in a simulated environment is straightforward. One can run the agent multiple times and take the average outcome. This is the widely used Monte Carlo method. Evaluating an agent in real-world scenarios is, however, much more challenging. This is because actually executing an agent in the real world can be very expensive; but building a high-fidelity simulator is often more challenging than solving the problem itself.

Recent *offline evaluation* methods take this challenge via performing policy evaluation with existing previously logged offline data, instead of new online data from agent-environment interaction. Offline methods typically rely on learning some other quantities, e.g., density ratio and action value function. However, those learning processes unavoidably suffer from various biases, resulting from, e.g., the misspecification of function class, the difficulty in optimizing nonconvex function, etc. As a result, offline methods rarely have evaluation accuracy guarantees without restrictive assumptions. Even worse, many offline methods still require a large amount of online data for model selection and thus do not really fulfill the promise to eliminate the use of online data. Offline model selection methods are being developed but rarely have model selection accuracy guarantees without restrictive assumptions. As a result, the Monte Carlo method using massive online data is still the most widely used policy evaluation method. The development and deployment of every RL algorithm implicitly or explicitly depend on Monte Carlo methods more or less. For example, when an RL researcher wants to plot the performance of an agent against training steps, the Monte Carlo method is almost always the first choice.

The Challenge

Given the dominance of Monte Carlo methods, I argue that it might be too ambitious at the moment to remove online data from our policy evaluation pipeline entirely. Reducing, instead of removing, the use of online data in Monte Carlo methods seems to be a more practical objective in the near future. Notably, the improved Monte Carlo methods should still be unbiased, or at least consistent, because it is the unbiasedness that makes Monte Carlo methods the golden standard in policy evaluation.

Efforts have been made to improve the data efficiency of Monte Carlo methods in RL, including Zinkevich et al. (2006); Hanna et al. (2017); Mukherjee, Hanna, and Nowak (2022); Zhong et al. (2022). Offline data is, however, not fully exploited to improve Monte Carlo methods yet, despite that offline data contains rich information about the environment. I, therefore, argue that the next important challenge in policy evaluation is to **reduce Monte Carlo methods' use of online data via information extracted from offline data while maintaining the unbiasedness of Monte Carlo methods.** Liu and Zhang (2023) shed light on this challenge via learning a provably variance-reducing behavior policy from offline data.

References

Hanna, J. P.; Thomas, P. S.; Stone, P.; and Niekum, S. 2017. Data-efficient policy evaluation through behavior policy search. In *Proceedings of the International Conference on Machine Learning*.

Liu, S.; and Zhang, S. 2023. Improving Online Monte Carlo Evaluation with Offline Data. *arXiv preprint*.

Mukherjee, S.; Hanna, J. P.; and Nowak, R. D. 2022. ReVar: Strengthening policy evaluation via reduced variance sampling. In *Proceedings of the Conference in Uncertainty in Artificial Intelligence*.

Zhong, R.; Hanna, J. P.; Schäfer, L.; and Albrecht, S. V. 2022. Robust On-Policy Data Collection for Data-Efficient Policy Evaluation. In *Advances in Neural Information Processing Systems*.

Zinkevich, M.; Bowling, M.; Bard, N.; Kan, M.; and Billings, D. 2006. Optimal unbiased estimators for evaluating agent performance. In *Proceedings of the AAAI Conference on Artificial Intelligence*.