Scaling Offline Evaluation of Reinforcement Learning Agents through Abstraction

Josiah P. Hanna

The University of Wisconsin – Madison
Madison, WI, U.S.A.
jphanna@cs.wisc.edu

Validating RL-Trained Policies

Interest in applying reinforcement learning (RL) has exploded as a promising way to learn decision-making policies in domains where it is difficult to manually pre-specify optimal actions. In practice, before a policy learned with RL is widely deployed – and its decisions have real-life consequences – we must be able to evaluate the expected outcome of letting the policy make those decisions. Evaluation is particularly critical when stakes are high, for example, when using RL to control expensive robots or drive cars at high speeds.

My research vision is to enable RL in challenging applications. As part of this vision, a central thrust of my work is to enable RL practitioners to deploy trained policies with the confidence that learned decision-making will represent improved performance in a variety of application domains. In this talk, I will present recent work from my group that aims to create tools for the accurate evaluation of RL-trained policies without real-world testing. Specifically, this work focuses on the *policy evaluation* problem in RL in which we are given a fixed *evaluation policy* and asked to estimate the expected cumulative reward that the evaluation policy would obtain if it were ran in the target deployment environment.

Scalable Offline Evaluation through Abstraction Recently, there has been much interest in methods for *offline policy evaluation* (OPE) as a means to perform policy evaluation without collecting any new data in a domain (Li 2019). Such methods use datasets of previously collected states, actions, and rewards to estimate the expected return that would be seen if the evaluation policy was deployed. Unfortunately, state-of-the-art OPE methods fall short of their promise – without large data sets of past decisions and their outcomes, current methods often fail to accurately evaluate untested policies in real-world settings.

To overcome this limitation, one of my students and I have developed new OPE methods that leverage the notion of a state abstraction function from the RL literature to scale OPE to domains where it has previously been unreliable. The key insight of this line of work is that, in OPE, the fundamental task is to adjust for differences between the evaluation policy and the policy (or policies) that generated the data used for evaluation. A limitation of prior OPE methods is that they try to adjust for *all* differences between past policies and the untested policy even *when such differences are irrelevant to evaluating policy performance*. In the presence of many irrelevant differences, state-of-the-art OPE methods can fail to return accurate estimates.

In this talk, I will first present our initial work in this direction where we showed how a *given state abstraction* could lead to more accurate OPE estimates both in theory and in practice (Pavse and Hanna 2023a). Here, the state abstraction provides a mechanism for a domain expert to specify state variables which are expected to be irrelevant and need not be adjusted for in OPE. I will then describe follow-up work that leveraged the idea of a *learned state representation* to further realize effective OPE in challenging and high-dimensional RL domains (Pavse and Hanna 2023b). This follow-up work enables irrelevant differences between states to be learned from data when a domain expert cannot identify them a priori. These two works serve as first steps in a direction that I envision will help realize the full potential of OPE for the safe and confident deployment of RL-trained policies. Finally, I will present directions for promising future work that aim to further leverage state and action abstraction to scale OPE to real world settings such as robotics, healthcare, and inventory control. These directions will create further scalable OPE methods and thus drastically increase practitioner confidence in deploying RL.

References

Li, L. 2019. A perspective on off-policy evaluation in reinforcement learning. *Frontiers of Computer Science*, 13(5): 911–912.

Pavse, B.; and Hanna, J. P. 2023a. Scaling Marginalized Importance Sampling to High-Dimensional State-Spaces via State Abstraction. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI)*.

Pavse, B. S.; and Hanna, J. P. 2023b. State-Action Similarity-Based Representations for Off-Policy Evaluation. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*.

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.