Offline Reinforcement Learning: Fundamental Barriers for Value Function Approximation

Dylan J. Foster DYLANFOSTER@MICROSOFT.COM
Akshay Krishnamurthy AKSHAYKR@MICROSOFT.COM
Microsoft Research

David Simchi-Levi DSLEVI@MIT.EDU
Yunzong Xu[*] YXU@MIT.EDU
Massachusetts Institute of Technology

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

We consider the offline reinforcement learning problem, where the aim is to learn a decision making policy from logged data. Offline RL—particularly when coupled with (value) function approximation to allow for generalization in large or continuous state spaces—is becoming increasingly relevant in practice, because it avoids costly and time-consuming online data collection and is well suited to safety-critical domains. Existing sample complexity guarantees for offline value function approximation methods typically require both (1) distributional assumptions (i.e., good coverage) and (2) representational assumptions (i.e., ability to represent some or all Q-value functions) stronger than what is required for supervised learning. However, the necessity of these conditions and the fundamental limits of offline RL are not well understood in spite of decades of research. This led Chen and Jiang (2019) to conjecture that *concentrability* (the most standard notion of coverage) and *realizability* (the weakest representation condition) alone are not sufficient for sample-efficient offline RL. We resolve this conjecture in the positive by proving that in general, even if both concentrability and realizability are satisfied, any algorithm requires sample complexity either polynomial in the size of the state space or exponential in other parameters to learn a non-trivial policy.

Our results show that sample-efficient offline reinforcement learning requires either restrictive coverage conditions or representation conditions that go beyond supervised learning, and highlight a phenomenon called *over-coverage* which serves as a fundamental barrier for offline value function approximation methods. A consequence of our results for reinforcement learning with linear function approximation is that the separation between online and offline RL can be *arbitrarily large*, even in constant dimension.

Keywords: Reinforcement learning, offline policy optimization/evaluation, function approximation, approximate dynamic programming, Markov decision process, sample complexity, fundamental limit

[*] Extended abstract. Full version appears as arXiv:2111.10919, v2. Part of this work was done while Y. Xu was an intern at Microsoft Research.