Towards Optimal Offline Reinforcement Learning

Daniel Kuhn

College of Management of Technology at EPFL, Switzerland
daniel.kuhn [at] epfl [dot] ch

We study offline reinforcement learning problems with a long-run average reward objective. The state-action pairs generated by any fixed behavioral policy thus follow a Markov chain, and the empirical state-action-next-state distribution satisfies a large deviations principle. We use the rate function of this large deviations principle to construct an uncertainty set for the unknown true state-action-next-state distribution. We also construct a distribution shift transformation that maps any distribution in this uncertainty set to a state-action-next-state distribution of the Markov chain generated by a fixed evaluation policy, which differs from the unknown behavioral policy. We prove that the worst-case average reward of the evaluation policy with respect to all distributions in the shifted uncertainty set provides the least conservative estimator for the average reward under the unknown true distribution. The emerging robust optimization problem can be viewed as a robust Markov decision process with a non-rectangular uncertainty set, and we develop an efficient policy gradient algorithm to solve this problem.