ID de Contribution: **288** Type: **Contributed talk**

Towards Optimal Offline Reinforcement Learning

mardi 29 juillet 2025 11:33 (24 minutes)

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 may differ 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, in a rigorous statistical sense, the least conservative estimator for the average reward under the unknown true distribution. This guarantee is available even if one has only access to one single trajectory of serially correlated state-action pairs. The emerging robust optimization problem can be viewed as a robust Markov decision process with a non-rectangular uncertainty set. We adapt an efficient policy gradient algorithm to solve this problem. Numerical experiments show that our methods compare favorably against state-of-the-art methods.

Author: LI, Mengmeng (EPFL)

Co-auteurs: KUHN, Daniel (EPFL); SUTTER, Tobias (University of Konstanz)

Orateur: LI, Mengmeng (EPFL)

Classification de Session: Mini-symposium