Fast Rates for the Regret of Offline Reinforcement Learning (Extended Abstract)

Yichun Hu YH767@CORNELL.EDU
Nathan Kallus KALLUS@CORNELL.EDU
Masatoshi Uehara MU223@CORNELL.EDU
Cornell University

Editors: Mikhail Belkin and Samory Kpotufe

Offline reinforcement learning (RL) is the problem of learning a reward-maximizing policy in an unknown Markov decision process (MDP) from data generated by running a fixed policy in the same MDP. The problem is particularly relevant in applications where exploration is limited but observational data plentiful. Medicine is one such example: ethical, safety, and operational considerations limit both the application of unproven or random policies and the running of online-updating algorithms, while at the same time rich electronic health records are collected en-masse.

A variety of methods have been proposed for offline RL including fitted Q-iteration (FQI), fitted policy iteration, modified Bellman Residual Minimization (BRM), SBEED, and MABO, among others. For all of these, the regret (value suboptimality) bounds obtained are $O(1/\sqrt{n})$, where n is the number of observed transition data. However, in practice, the regret convergence can actually be *much* faster. For example, we provide a linear-MDP simulation experiment where FQI empirically exhibits an apparent regret convergence rate of $O(1/n)$.

In this paper, we tightly characterize this phenomenon by theoretically establishing fast rates for the regret convergence of value-based offline RL methods, which directly estimate the optimal quality function, Q^*. These rates leverage the specific noise level of a given problem instance, expressed as the density near zero of the suboptimality of the second-best action (if any), also known as a *margin condition*. RL instances generally satisfy *some* instance-specific nontrivial margin condition. We moreover show that in the linear and tabular cases, we generally have quite strong margin conditions. We show that policies that are greedy with respect to good estimates of Q^* enjoy a regret bounded by the pointwise estimation error raised to a power *larger* than one, thus speeding up convergence for the downstream decision-making task. This analysis can be applied to any value-based offline RL method that has pointwise convergence guarantees for estimating Q^*. As specific examples, we establish that we can achieve such pointwise error bounds for the linear case using FQI and modified BRM (differently from existing analyses of their average error). Together, this means that, under the standard assumptions needed for FQI and modified BRM, *i.e.*, closedneess under Bellman operators (completeness) and sufficient feature coverage, linear FQI and modified BRM generally achieve regret of order $O(1/n)$ in linear MDPs. Technically, our analysis melds ideas from fast-rate analysis of classification with the theoretical analysis of RL.[1]

[1] Extended abstract. Full version appears as [arXiv:2102.00479, v1].