Learning to Stop with Surprisingly Few Samples

Daniel Russo DJR2174@GSB.COLUMBIA.EDU
Graduate School of Business, Columbia University
Assaf Zeevi ASSAF@GSB.COLUMBIA.EDU
Graduate School of Business, Columbia University
Tianyi Zhang TZ2376@GSB.COLUMBIA.EDU
Graduate School of Business, Columbia University

Editors: Mikhail Belkin and Samory Kpotufe

We consider a discounted infinite horizon optimal stopping problem. A sequence of i.i.d. random variables $X_1, \ldots, X_n$ are revealed sequentially. Fixing a discount factor $\gamma \in (0, 1)$, the player seeks to solve

$$V^* = \sup_{1 \le \tau \le n} \mathbb{E}_F \left[\gamma^\tau X_\tau\right]. \tag{1}$$

If the underlying distribution is known a priori, the solution of this problem is obtained via dynamic programming (DP) and is given by a well known threshold rule. When information on this distribution is lacking, the challenge is to leverage this structural property with a suitable learning algorithm. Goldenshluger and Zeevi (2021) proposes a rank-based policy for a finite horizon problem which is proven to be asymptotically optimal relative to the full information DP solution. An open question is whether simpler families of policies might yield competitive performance.

A natural (though naive) approach is "explore-then-exploit," whereby the unknown distribution or its parameters are estimated over an initial exploration phase, and this estimate is then used in the DP to determine actions over the residual exploitation phase. While common wisdom suggests that solutions to the Bellman equation can be quite sensitive even to "small" estimation errors, we show that collecting an amount of data that scales only logarithmically in the "effective horizon" is sufficient to support near optimal solutions in the case of optimal stopping. In particular, for a reasonably broad parametric class of distributions, we show that an exploration phase that collects only on the order of $(\log(1/(1-\gamma)))^2$ observations from F suffices to make such "plug in" approach near optimal. This threshold is sharp, in the sense that any number of samples which is of lower order is catastrophic for the decision-maker, with guaranteed loss of at least half of the optimal value. Surprisingly, the length of the exploration horizon required to support near-optimal performance is smaller in problems where the underlying distribution have tails that decrease more slowly. This is especially pronounced when the tails are heavy where a *single sample* exploration phase suffices. The latter clearly indicates how common intuition pertaining to the negative effects of heavy tails may be false when estimation goals are pursued concurrent to decision making (optimization) objectives.

The paper offers a detailed "case-study" that reveals rich and unexpected behavior. As such, it motivates further investigation and may serve as a basis for studying sample complexity of learning in structured classes of dynamic optimization problems. In particular, it may be of broader intellectual relevance in studying the impact of time horizon on sample complexity of both online and offline reinforcement learning (e.g., Zhang et al. (2020)).

. Extended abstract. For full version see https://arxiv.org/abs/2102.10025

References

A Goldenshluger and A Zeevi. Optimal stopping of a random sequence with unknown distribution. *Mathematics of Operations Research (to appear)*, 2021.

Zihan Zhang, Xiangyang Ji, and Simon S Du. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. *arXiv preprint arXiv:2009.13503*, 2020.