Reliable Offline RL in POMDPs: Safe Policy Improvement for POMDPs via Finite-State Controllers

Thiago D. Simão – PostDoc at Radboud University

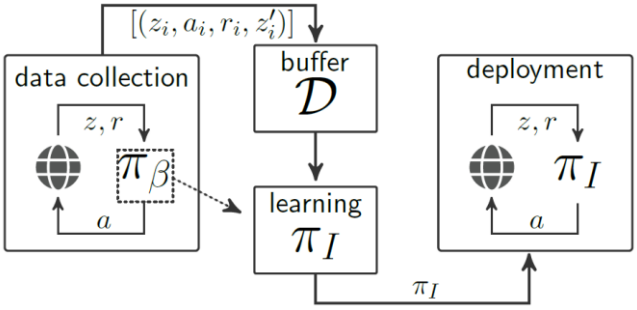
We study the safe policy improvement (SPI) problem for partially observable Markov decision processes (POMDPs). SPI is an offline reinforcement learning (RL) problem that assumes access to (1) historical data about an environment, and (2) the so-called behaviour policy that previously generated this data by interacting with the environment. The underlying SPI method constrains the policy-space according to the available data, such that the newly computed policy only differs from the behaviour policy when sufficient data was available. We show that this new policy, converted into a new FSC for the (unknown) POMDP, outperforms the behaviour policy with high probability.

Introduction

Reinforcement learning (RL) is a standard approach to solve sequential decision-making problems when the environment dynamics are unknown [1]. Typically, an RL agent interacts with the environment and optimizes its behavior according to environment's feedback. However, in offline RL (Levine et al., 2020), the RL agent receives a fixed dataset of past interactions between a behavior policy and the environment and derives a new policy with no direct interaction with the environment. One of the challenges in offline RL is to ensure that the new policy outperforms the behavior policy. This problem is called *safe policy improvement* (SPI; Thomas et al., 2015a). Most approaches to SPI assume fully observable environments, see for instance (Laroche et al., 2019).

Methods and approach

The restriction to full observability limits the applicability of SPI, as most real-world problems are *partially observable*, due to, for instance, noisy sensors. Partially observable Markov decision processes (POMDPs) are the standard model for decision-making problems under partial observability (Kaelbling et al., 1998). So far, SPI for POMDPs was only studied for



memoryless policies (Thomas et al., 2015b). However, POMDP policies often require a notion of memory. In general, optimal policies for POMDPs with infinite horizons require infinite memory, rendering this problem undecidable. Nevertheless, finite memory can approximate the optimal policy and are often used in practice for being more explainable. Policies with finite memory may take the form of finite-state controllers (FSCs).

Proposed solution

We contribute a novel SPI approach for POMDPs. First, to account for the inherent memory requirement in partially observable domains, we present the behavior policy as an FSC. To create a tractable method, we assume that there exists a finite-memory policy for the

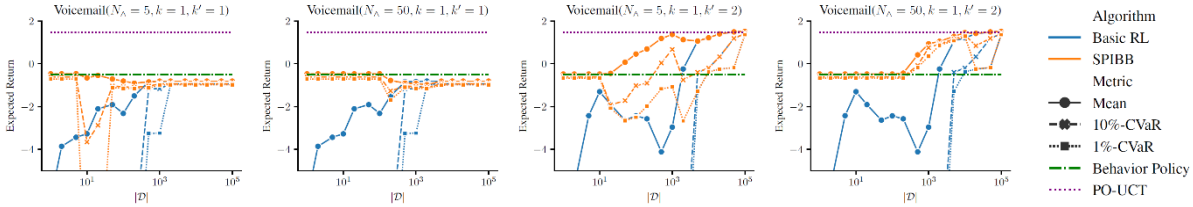


Figure 1. Policy improvement on the Voicemail environment for datasets collected with a memoryless policy (k = 1), varying the hyperparameters pairs column-wise. The plots show the mean (solid line), 10%-CVaR (dashed line) and 1%-CVaR (dotted line). The performance of the behavior policy is shown in green (dash-dotted line).

POMDP that is optimal. This assumption allows us to cast the POMDP as an equivalent, fully observable, history MDP that is finite, instead of the standard infinite-history MDP. We are then able to reliably estimate the transition and reward models of this finite-history MDP from the available data. We employ the algorithm *safe policy improvement with baseline bootstrapping* (SPIBB; Laroche et al., 2019). In particular, we compute an improved policy that outperforms the behavior policy with high probability.

Potential Applications

This setting captures multiple real-world applications, such as predictive maintenance, conservation of endangered species, and management of invasive species. We may, for instance, have data from the degradation process of a certain asset, which includes logs of inspections and maintenance that were performed according to a fixed schedule (represented, for instance, as a finite-state controller). Once we acquire a new asset, we can formalize the optimization problem with offline RL to compute a new schedule, using the original schedule as a behavior policy.

Conclusion

We presented a new approach to safe policy improvement in POMDPs. Our experiments show the applicability of the approach, even in cases where finite-history is not sufficient to obtain optimal results.

References

Richard Sutton and Andrew Barto. Reinforcement Learning - An introduction. MIT Press, 1998.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems. arXiv:2005.01643, 2020.

Philip S. Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In ICML, 2015a.

Romain Laroche, Paul Trichelair, and Remi Tachet des Combes. Safe Policy Improvement with Baseline Bootstrapping. In ICML, 2019.

Leslie Kaelbling, Michael Littman, and Anthony Cassandra. Planning and acting in partially observable stochastic domains. Artif. Intell., 1998.

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In AAAI, 2015b.

Facts

Researchers	Thiago D. Simão * Marnix Suilen Nils Jansen
Academic partners	Radboud University
	* thiago.simao@ru.nl

https://arxiv.org/abs/2301.04939