Offline Policy Evaluation for Reinforcement Learning with Adaptively Collected Data

October 3, 2022

Abstract

Offline RL is an important step towards making data-hungry RL algorithms more widely usable in the real world, but conventional assumptions on the distribution of logging data do not apply in some key real-world scenarios. In particular, it is unrealistic to assume that RL practitioners will have access to sets of trajectories that simultaneously are mutually independent and explore well. We propose two natural ways to relax these assumptions: by allowing the data to be distributed according to different logging policies independently, and by allowing logging policies to depend on past trajectories. We discuss Offline Policy Evaluation (OPE) in these settings, analyzing the performance of a model-based OPE estimator when the MDP is tabular.

1 Introduction

The central goal of Reinforcment Learning (RL) is to understand how to act optimally in a stochastic dynamical system, where our actions influence the system's state (typically, a Markov Decision Process (MDP)). *Online* algorithms, which gather data about the MDP by taking actions and observing outcomes, can be very successful at optimizing behavior for a reward function. Unfortunately, in many scenarios where we might want to engage in this kind of sequential learning, it would be too dangerous to run algorithms that explore randomly. Even in (comparatively) low-stakes scenarios, such as advertising, adopting elementary RL algorithms would mean throwing away vast reserves of data gathered from previous strategies. Both of these problems would be partially addressed by practical methods for *offline* learning, which seeks to solve the control problem using existing data.

The offline reinforcement learning problem can be split into two goals: Offline Policy Evaluation (OPE), and Offline Learning (OL). Efficient OPE is often a stepping-stone on the way to OL.

In OPE, we seek to evaluate a policy, π, using data collected from some distinct logging policy, μ. More formally, let $\mathcal{M}$ be an MDP, and μ be a logging-policy. Given a dataset, $\mathcal{D}$, consisting of n trajectories distributed independently according to μ and $\mathcal{M}$, we wish to construct an estimator $\hat{v}^\pi$ of v^π. Performance guarantees on OPE estimators can take the form of upper bounds on $MSE(\hat{v}^\pi)$ (Yin and Wang, 2020), or (stronger) high probability bounds of the form $|v^\pi - \hat{v}^\pi| < \epsilon$ (Yin et al., 2020).

On the other hand, the OL problem asks us to use $\mathcal{D}$ to find a good policy $\pi \in \Pi$, where Π is some class of policies. We desire that this π satisfy a condition like $|v^* - v^\pi| < \epsilon$ w.h.p., where v^* is the value of the optimal policy π^* in the policy class Π.

Practically, OL would allow RL practitioners to exercise more control over the gathering of training data. However, in order for $\mathcal{D}$ to be a rich enough dataset to learn from, strong assumptions need to be made about the exploratory properties of μ. This is in addition to the already strong

assumption that the set of trajectories is i.i.d. In Section 3, we discuss the problems that such assumptions pose and propose ways of weakening them.

2 Notation

Let $\Delta(\mathcal{X})$ be the set of all PMFs over $\mathcal{X}$, for $|\mathcal{X}| < \infty$. Let $[H] := \{1, ..., H\}$

A Tabular, Finite-Horizon Markov Decision Process is a tuple $(\mathcal{S}, \mathcal{A}, r, P, d_1, H)$, where $\mathcal{S}$ is the state space ($|\mathcal{S}| =: S$), and $\mathcal{A}$ is the action space ($|\mathcal{A}| =: A$). Its dynamics are governed by a nonstationary transition kernel, $P = \{P_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})\}_{h=1}^H$, where $P_h(s'|s, a)$ is the probability of transitioning to state $s' \in \mathcal{S}$ after taking action $a \in \mathcal{A}$ from state $s \in \mathcal{S}$ at time $h \in [H]$. r is a collection of reward functions $\{r_h : \mathcal{S} \times \mathcal{A} \to [-1, 1]\}_{h=1}^H$. Finally, $d_1 \in \Delta(\mathcal{S})$ is the initial state distribution of the MDP and H is the horizon.

A policy, π, is a collection of maps, $\{\pi_h : \mathcal{S} \to \Delta(\mathcal{A})\}_{h=1}^H$.

Running a policy on an MDP will yield a trajectory $\tau_i \in (\mathcal{S} \times \mathcal{A} \times [-1, 1])^H$. Together, the policy and MDP induce a distribution over trajectories, as well as a Markov Chain with transitions notated as $P_h^\pi(s'|s) := \sum_a P_h(s'|s, a)\pi_h(a|s)$.

In a set of trajectories $\{\tau_i\}_{i=1}^n$, we define $n_{h,s,a}$ to be the number of visitations to (s, a) at timestep h.

$v^\pi := \mathbb{E}_\pi[\sum_{i=1}^H r_i | s_1 \sim d_1]$ is the value of the policy π, where the expectation is over the π-induced distribution over trajectories.

$d_h^\pi(s, a)$ is defined to be the probability of (s_h, a_h) occurring at time step h in a trajectory distributed according to π.

3 Problem Formulation and Motivation

Naturally, both OPE and OL are hopeless if the logging-policy does not explore well. If the logger, μ, does not visit a state that the target policy, π, visits very often, we will not be able to form an accurate estimate of the target value v^π when doing OPE. In OL, this issue is compounded by the fact that missing *any* state may correspond to missing high-value outcomes. Thus, upper bounds on the performance of OPE or OL algorithms are given in terms of an exploration parameter, like

$$d_m = \min_{h,s,a:d_h^\pi(s,a)>0} d_h^\mu(s, a) \tag{1}$$

Estimators that closely match lower bounds on relevant metrics have been established in existing work, such as Duan and Wang (2020). However, the practicability of such bounds has been challenged. Xiao et al. (2021) point out that it is difficult in practice to find a logging policy with a reasonable exploration parameter. In what they consider a more realistic, "tabula rasa" case (where the logging policy is chosen without knowledge of the MDP), they show a sample complexity exponential in H and S to be necessary in offline learning.

Besides this limitation, we note that the motivating application of learning from existing, human-generated data remains unfulfilled in the conventional OPE setting. We would not expect these data to be identically distributed, or even independent (as the data collected in trajectory j almost certainly influences future policies $\mu^{j+1}, ...$).

Thus motivated, we augment our formulation of the OPE problem to more realistically accommodate intelligent choices of logging-policy. We consider this to be middle ground between the sanguine assumptions on d_m common to Yin and Wang (2020); Yin et al. (2020); Duan and Wang (2020), and the assumption of total ignorance found in Xiao et al. (2021). To this end, this paper studies two settings:

1. OPE on $\mathcal{D} = \{\tau_i \sim \mu^i\}_{i=1}^n$, where $\mu^1, ...\mu^n$ are distinct policies chosen in advance, and the trajectories unfold independently. We call this the Non-Adaptive OPE (NOPE).

2. OPE on $\mathcal{D} = \{\tau_i \sim \mu^i\}_{i=1}^n$, where $\mu^1, ...\mu^n$ are chosen adaptively. That is, μ^i may depend on the trajectories $\tau_1, ...\tau_{i-1}$. We note, however, that the trajectories $\{\tau_i\}_i$ are independent conditioned on $\{\mu^i\}_i$. We call this Adaptive OPE (AOPE).

NOPE allows us to hedge our bets with respect to the logging policy we choose. In particular, when we derive a bound on the MSE of the model-based estimator from (Yin and Wang, 2020), it will be in terms of the minimum of the average state-action occupancy:

$$\bar{d}_m := \frac{1}{n} \min_{h,s,a} \sum_{i=1}^{n} d_h^{\mu^i}(s,a) \tag{2}$$

If there is some pre-existing knowledge on the MDP, $\mathcal{M}$, it may be easier to propose n logging policies with (2) bounded away from zero than a single logging policy with (1) bounded away from zero.

NOPE does not adequately address concerns over finding good logging policies, but rather dilutes the problem to that of finding a suite of policies that is good on average. Furthermore, it only eliminates assumptions on identical distribution, and sets aside issues raised earlier over dependence between trajectories.

Both of these problems are better addressed by AOPE. When logging policies can be tuned according to previous trajectories, there is scope for starting from "tabula rasa", and iteratively refining the logging policy as we learn about the MDP. In other words, the logger can leverage online exploration techniques. Furthermore, by allowing arbitrary statistical dependence on previous trajectories, AOPE addresses the key scenario of learning from intradependent, manually-collected datasets.

4 Method

We consider the estimator of v^π studied in (Yin and Wang, 2020). This boils down to computing the value of a policy under the approximate MDP defined by $(\mathcal{S}, \mathcal{A}, \hat{P}, \hat{r}, \hat{d}_1)$, with the estimators $\hat{P}$, $\hat{r}$ and $\hat{d}_1$ defined below.

That is, if $\mathcal{D} = \{\tau_1, ...\tau_n\}$, and $\tau_i = (s_1^i, a_1^i, r_1^i, ...s_H^i, a_H^i, r_H^i)$, we use plug-in estimates

$$\hat{P}_h(s'|s,a) = \frac{n_{h,s,a,s'}}{n_{h,s,a}} = \frac{1}{n_{h,s,a}} \sum_i \mathbb{1}_{\{s_h^i = s, a_h^i = a, s_{h+1}^i = s'\}} \qquad \hat{r}_h(s,a) = \frac{1}{n_{h,s,a}} \sum_{k=1}^{n} r_h^k \mathbb{1}_{\{s_h^k = s, a_h^k = a\}}$$

subject to these quantities being well-defined ($n_{h,s,a} \neq 0$). If $n_{h,s,a} = 0$, we can define them to be 0.

We also define $\hat{d}_1 :=: \hat{d}_1^\pi := \frac{1}{n} \sum_{i=1}^n e_{s_1^i}$ to be the plug-in estimate of d_1 computed from $\mathcal{D}$ (where e_j is the jth standard basis vector in $\mathbb{R}^S$).

We then let:

$$\hat{P}_h^\pi(s'|s) = \sum_a \pi_h(a|s) \hat{P}_h(s'|s,a) \qquad \hat{r}_h^\pi(s) = \sum_a \pi_h(a|s) \hat{r}_h(s,a)$$

and iteratively compute $\hat{d}_h^\pi := \hat{P}_h^\pi \hat{d}_{h-1}^\pi$ for $h = 1, ...H$.

Finally, we form the estimate

$$\hat{v}^\pi = \sum_{h=1}^{H} \langle \hat{d}_h^\pi, \hat{r}_h^\pi \rangle$$

5 Current Results

By following the proof of Theorem 1 in Yin and Wang (2020), and making some mild modifications, we recover a bound of the MSE of $\hat{v}^\pi$ in the Non-Adaptive OPE setting.

Theorem 1. *[MSE performance of $\hat{v}^\pi$ in NOPE setting] Suppose $\mathcal{D}$ is a dataset conforming to NOPE, and $\hat{v}^\pi$ is formed using this dataset. Let $\bar{d}_m$ be as defined in (2). Let $\tau_s = \max_{h,s,a} \frac{d_h^\pi(s,a)}{\frac{1}{n}\sum_i d_h^{\mu^i}(s,a)}$. Let $\tau_a = \max_{h,s,a} \frac{\pi_h(a|s)}{\frac{1}{n}\sum_i \mu_h^i(a|s)}$. Then if $n > \frac{16\log n}{\bar{d}_m}$ and $n > \frac{4H\tau_a\tau_s}{\min_{h,s}\max\{d_h^\pi(s), \frac{1}{n}\sum_i d_h^{\mu^i}(s)\}}$ we have:*

$$MSE(\hat{v}^\pi) \leq \left(1 + \sqrt{\frac{16\log n}{n\bar{d}_m}}\right)\frac{1}{n}\sum_{h,s,a}\frac{d_h^\pi(s)^2\pi(a|s)^2}{\sum d_h^{\mu^i}(s,a)}\mathrm{Var}[r_h^{(1)} + V^\pi(s_{h+1}^{(1)})|s_h^{(1)} = s, a_h^{(1)} = a]$$

$$+O(\tau_a^2\tau_s H^3/n^2\bar{d}_m)$$

As a corollary, consider a "quasi-adaptive" data collection process, where each logging-policy, μ^i, is run twice, generating i.i.d. τ_i and τ_i'. Suppose future logging-policies $\mu^{j>i}$ are chosen by some algorithm $\mathcal{E}$ depending on τ_i but not τ_i'. We can use the same $\hat{v}^\pi$-estimator to perform OPE with $\mathcal{D}' = \{\tau_i'\}$, as long as the estimator doesn't touch $\mathcal{D} = \{\tau_i\}$. If we assume that average exploration is sufficient w.h.p. over the execution of $\mathcal{E}$, we can the bound MSE in this quasi-adaptive case using Theorem 1, the fact that the $\{\tau_i'\}$ are mutually independent conditioned on $\{\mu^i\}$, and the tower rule.

Corollary 2. *Let $\mathcal{E}$ be the algorithm described in the paragraph above. Assume that with high probability ($\geq 1-\delta$), the policies $\mu^1, ...\mu^n$ generated by $\mathcal{E}$ satisfy $\frac{1}{N}\sum_i d_h^{\mu^i}(s,a) \geq \bar{d}_m$ for all $s \in S$ and $a \in A$, and for some $\bar{d}_m$. Then*

$$MSE(\hat{v}^\pi) \leq (1-\delta)(*) + H^2\delta$$

where $(*)$ is the bound on the MSE of the estimator in the non-adaptive case from Theorem 1.

We now turn our attention towards quantifying $\hat{v}^\pi$'s performance on AOPE. We first describe a high-probability, uniform error bound in terms of the number of visitations to each (s_h, a_h) tuple.

Theorem 3 (High-probability uniform bound on estimation error in AOPE)**.** *Suppose $\mathcal{D}$ is a dataset conforming to AOPE, and $\hat{v}^\pi$ is formed using this dataset. Then, with probability at least $1-\delta$, the following holds for all policies π.*

$$|\hat{v}^\pi - v^\pi| \leq K\sum_{h=1}^{H}\sum_{s,a} H d_h^\pi(s,a)\sqrt{\frac{S\log\frac{HSAn}{\delta}}{n_{h,s,a}}}$$

where $n_{h,s,a}$ is the number of occurrences of (s_h, a_h) in $\mathcal{D}$, and K is an absolute constant.

This translates to the following bound, which underperforms the upper bound for deterministic policies in the non-adaptive case in Yin et al. (2020) by a factor of $\sqrt{H}$.

Corollary 4 (High-probability uniform bound on estimation error in AOPE)**.** *Suppose that $\mathcal{D}$, $\hat{v}^\pi$ are as in Theorem 3. Then with probability $1-\delta$, we have that*

$$\sup_\pi |\hat{v}^\pi - v^\pi| \leq O(H^2\sqrt{\frac{S\log HSAn/\delta}{n\bar{d}_m}})$$

The proof of Theorem 3 follows by a simulation lemma-type expansion of the error, which leads to a dominant term of the form $\sum_h \mathbb{E}_{s_h,a_h\sim\pi,\mathcal{M}}[(\hat{P}_{h+1}(\cdot|s_h,a_h) - P_{h+1}(\cdot|s_h,a_h))^T \hat{V}^\pi_{h+1}]$, and smaller terms governed by $\hat{r}$ and $\hat{d}_1$. In order to get around the issue of dependence between trajectories, we cover all possible number of occurrences of each (s_h, a_h) across trajectories while applying concentration, leading to to the $HSAn$ term inside the logarithm.

We also give a high-probability pointwise bound, which is suboptimal by a factor of H. In the pointwise case, we are able to shave off a $\sqrt{S}$ in the asymptotically dominant term.

Theorem 5 (Pointwise bound on estimation error in AOPE)**.** *Fix a policy π, suppose $\mathcal{D}$ is a dataset conforming to AOPE, and $\hat{v}^\pi$ is formed using this dataset. Then with probability at least $1-\delta$, we have:*

$$|\hat{v}^\pi - v^\pi| \leq O(H^2\sqrt{\frac{\log HSAn/\delta}{n\bar{d}_m}} + \frac{H^2 S\frac{HSAn}{\delta}}{n\bar{d}_m})$$

Inspired by Azar et al. (2017), Theorem 5 is proved by applying concentration inequalities (with the same covering trick as Theorem 3) to $(\hat{P}_{h+1} - P_{h+1})V^\pi_{h+1}$ and $(\hat{P}_{h+1} - P_{h+1})(\hat{V}^\pi_{h+1} - V^\pi_{h+1})$ separately, instead of $(\hat{P}_{h+1} - P_{h+1})\hat{V}^\pi_{h+1}$. A factor of $\sqrt{H}$ could likely be saved by using Bernstein's inequality on $(\hat{P}_{h+1} - P_{h+1})V^\pi_{h+1}$ instead of Hoeffding, but we quote the latter.

Next Steps

We note that Theorem 3 is worse by a factor of $\sqrt{H}$ when compared to corresponding bounds for the non-adaptive setting (Yin et al., 2020). This is because we could not salvage the martingale structure leveraged across timesteps in their analysis. As future work, we intend to recover better bounds by more carefully analyzing the variance structure of the MDP, or else demonstrate (by means of a lower bound) that the estimator $\hat{v}^\pi$ is worse at AOPE than OPE. Similarly, we plan on tightening the bound provided in Theorem 5. We believe that theoretically clarifying the extent to which OPE methods carry over to the AOPE setting is a step towards making offline RL a yet more convincing candidate for real-world decision problems.

References

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning, 2017. URL `https://arxiv.org/abs/1703.05449`.

Yaqi Duan and Mengdi Wang. Minimax-optimal off-policy evaluation with linear function approximation. *CoRR*, abs/2002.09516, 2020. URL `https://arxiv.org/abs/2002.09516`.

Chenjun Xiao, Ilbin Lee, Bo Dai, Dale Schuurmans, and Csaba Szepesvari. The curse of passive data collection in batch reinforcement learning, 2021. URL `https://arxiv.org/abs/2106.09973`.

Tengyang Xie, Yifei Ma, and Yu-Xiang Wang. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling, 2019. URL `https://arxiv.org/abs/1906.03393`.

Ming Yin and Yu-Xiang Wang. Asymptotically efficient off-policy evaluation for tabular reinforcement learning. 2020. doi: 10.48550/ARXIV.2001.10742. URL `https://arxiv.org/abs/2001.10742`.

Ming Yin, Yu Bai, and Yu-Xiang Wang. Near-optimal provable uniform convergence in offline policy evaluation for reinforcement learning, 2020. URL https://arxiv.org/abs/2007.03760.