Offline Policy Evaluation for Reinforcement Learning with Adaptively Collected Data

Sunil Madhow   Dan Qiao   Ming Yin   Yu-Xiang Wang
UC Santa Barbara
{sunilmadhow,danqiao,ming_yin}@ucsb.edu
yuxiangw@cs.ucsb.edu

January 22, 2023

Abstract

Developing theoretical guarantees on the sample-complexity of offline RL methods is an important step towards making data-hungry RL algorithms practically viable. Currently, such results hinge on unrealistic assumptions about the data distribution – namely that it comprises a set of i.i.d. trajectories collected by a single logging policy. 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 believe the latter, “adaptive" regime better captures the settings in which Offline RL algorithms are already being gainfully applied. We discuss Offline Policy Evaluation (OPE) in these settings, analyzing the performance of a model-based OPE estimator when the MDP is tabular. Finally, we conduct simulations to empirically analyze the behavior of these estimators under adaptive and non-adaptive regimes.

1 Introduction

Offline RL, which seeks to perform standard RL tasks using a pre-existing dataset of interactions with an MDP, is a key frontier in the effort to make RL methods more widely applicable. The ability to incorporate existing data into RL algorithms is crucial in many promising application domains. In safety-critical areas, such as autonomous driving and medicine, the randomized exploration that characterizes online algorithms is not ethically tolerable. Even in lower-stakes applications, such as advertising, naively adopting online algorithms could mean throwing away vast reserves of previously-collected data. The development of efficient offline algorithms promises to broaden RL's applicability by allowing practitioners to exercise some much needed domain-specific control over the training process.

Given a dataset, $\mathcal{D}$, of interactions with an MDP, two tasks that we may hope to achieve in offline RL are Offline Policy Evaluation and Offline Learning. In Offline Policy Evaluation (OPE), we seek to estimate the value of a target policy π under $\mathcal{M}$. In Offline Learning (OL), the goal is to use $\mathcal{D}$ to find a good policy $\pi \in \Pi$ where Π is some policy class. In this paper, we restrict our attention to the easier of the two: OPE.

The question of how and when it is possible to perform OPE and OL given a specific dataset, $\mathcal{D}$, is the subject of a flourishing research movement. Clearly, in order for $\mathcal{D}$ to be a rich enough dataset to learn from, strong assumptions need to made about how well it explores the MDP. Excellent results have been derived assuming that $\mathcal{D}$ consists of i.i.d. trajectories distributed according to to some logging policy μ, where μ has good exploratory properties. However, it is difficult to justify the imposition of these assumptions on the dataset. Empirically, it is clear that the generation of a usable $\mathcal{D}$ is a significant bottleneck in offline RL. In practice, the gathering of useful datasets is best done by running adaptive exploration algorithms (see, for example, Lambert et al. (2022)'s use of "curiosity"). Even supposing that a good μ were handed to us, it is not likely that we would be willing to blindly run-it in a real-world situation. In the case of self-driving cars, *any* dataset will be rife with adaptivity due to human interventions and the need to iterate policies. If the theory of offline RL is to be relevant in real-world use-cases, it must extend to adaptively collected datasets.

Our object in this paper is to develop a systematic understanding of the role of adaptivity in offline RL. In Section 4, we discuss the problems posed by the i.i.d. assumption from a more rigorous perspective, and propose weaker settings for OPE. We believe the most fruitful of these to be Adaptive OPE (AOPE), where we allow each trajectory to be distributed according to a different logging policy, which may depend on previous data.

In addition to the motivating examples given above, here are some scenarios that AOPE covers and OPE does not are presented below.

1. The dataset $\mathcal{D}$ has been collected over a long period of time, during which unrecorded changes have been made to the policy. An example of this might be the learning outcomes of students on a changing online curriculum.

2. The dataset $\mathcal{D}$ was gathered by humans, and therefore influenced by a number of unobserved factors. For example, a doctor prescribing medicine may make a determination based on her conversation with the patient – a factor not recorded in any state variable.

3. The dataset $\mathcal{D}$ has been gathered by a low-regret algorithm, like UCB-VI. This dataset will have excellent exploratory properties, but is very intradependent.

If it can be demonstrated that existing OPE estimators extend to the AOPE setting with no performance drop, a new class of use-cases for Offline RL emerges. If AOPE is, in fact, verifiably harder than OPE, practitioners will know that to enjoy strong guarantees in Offline RL tasks, they will need to collect remarkably clean data. This paper represents our first steps towards resolving this key question. In Section 6, we explain how Yin et al. (2021)'s minimax-optimal results for both Uniform OPE and efficient learning lead directly to results for the adaptive case. We proceed to derive more evocative *instance dependent* bounds on the estimation error $|\hat{v^{\pi}} - v^{\pi}|$. Our instance-dependent bounds may significantly outperform the minimax-optimal bound on certain instances, but they fail to recover optimal *worst-case* rates.

In Section 7, we empirically study how adaptivity can improve or degrade the performance of our estimator.

2 Related Work

The OPE literature is vast. We do not attempt to provide a survey of the excellent body of work but instead refer readers to the recent work of Mou et al. (2022) and the references therein for a more comprehensive discussion. To the best of our knowledge, we are the first to study the problem of OPE under the adaptive data setting. Most existing work on OPE that we have seen makes the assumption that the data are collected iid from a single logging policy. The only exception is the work of Kallus et al. (2020) who studied OPE from multiple loggers, but they only considered the contextual bandits model and non-adaptive loggers, while we studied the RL with possibly adaptively chosen loggers.

OPE is also closely related to the average treatment effect estimation problem in the causal inference literature, but typically only only one-step decision is considered and the observational data are assumed to be iid.

3 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. Similarly, $V^{\pi}(s) := \mathbb{E}_{\pi}[\sum_{i=1}^{H} r_i | s_1 = s]$.

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

4 Problem Formulation and Motivation

Consider the standard setting where $\mathcal{D} = \{\tau_i \sim \mu\}_{i=1}^{n}$ is a collection of iid trajectories. 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, bounds on the performance of OPE or OL algorithms are given in terms of

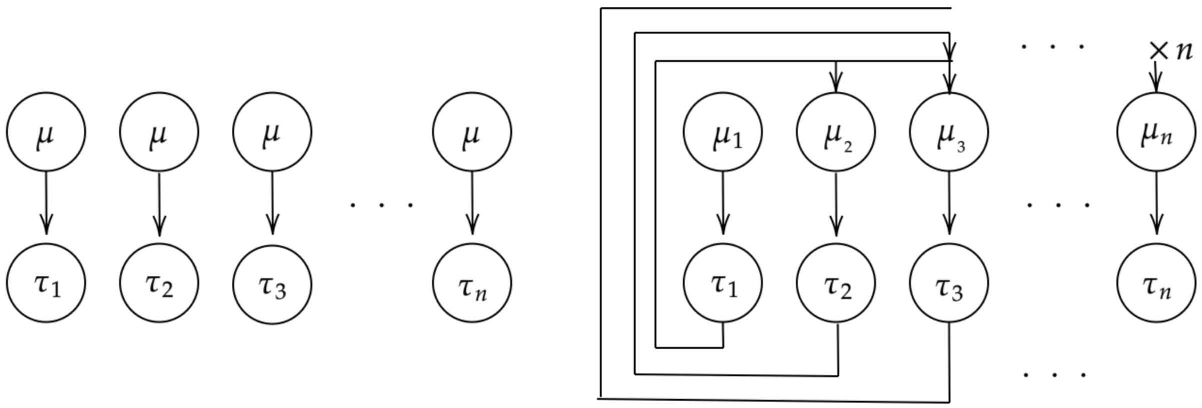

Figure 1: Non-adaptive regime (left) versus adaptive regime (right), depicted as a graphical model. We see that, in the adaptive regime, each policy depends on all previous trajectories. This induces dependence between the trajectories.

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 et al. (2020). However, the practicability of such bounds has been challenged. Xiao et al. (2022) 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 et al. (2021); Duan et al. (2020), and the assumption of total ignorance found in Xiao et al. (2022). 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 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 preexisting 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.

We conclude this section by noting that minimax bounds from the non-adaptive OPE setting can be easily recovered for AOPE in the following manner. For any dataset, $\mathcal{D}$, let $N := \min_{h,s,a} n_{h,s,a}$ be the number of occurrences of the least-observed (s_h, a_h) pair. If we consider a revised dataset, $\mathcal{D}'$, that keeps only the first N transitions out of (s_h, a_h) for all s_h, a_h, we see that all transitions are now independent conditioned on N. Thus, the problem reduces to a generative-model type setting. In particular, Theorem 3.7 in Yin et al. (2021) implies minimax-optimal offline learning for the adaptive case. However, we do not like throwing data away in this manner, and conjecture that it should not be necessary to do so. It would be preferable to obtain bounds that adapt to the quantities $\{n_{h,s,a}\}_{h,s,a}$ and the features of the MDP. To this end, we explore the extent to which *instance-dependent* bounds on estimation error can be recovered in the adaptive setting.

5 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:

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

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

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

6 Theoretical Results

We warm up by generalizing Yin and Wang (2020)'s bound on the MSE of $\hat{v}^\pi$ to NOPE. 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:*

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

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

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 $\mathcal{D}_{shadow} = \{\tau_i'\}$, 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.*

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

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 worst-case bound, which underperforms the minimax-optimal bound (over deterministic policies) implied by Yin et al. (2021) 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. Suppose that, with probability $\ge 1-\delta/2$, the logging process is such that $n_{h,s,a} \ge n\bar{d}_m$ for all h, s, a. Then with probability $1-\delta$, we have that*

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

The proof of Theorem 3 follows by a simulation lemma-type expansion of the error, which leads to a dominant term of the form $\sum_h \mathbb{E}_{s_h,a_h \sim \pi,\mathcal{M}}[(\hat{P}_{h+1}(\cdot|s_h,a_h) - P_{h+1}(\cdot|s_h,a_h))^T \hat{V}_{h+1}^\pi]$, 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, instance-dependent, *pointwise* bound, which is suboptimal by a factor of $\sqrt{H}$ when translated into a worst-case bound. In the pointwise case, we are able to shave off a $\sqrt{S}$ in the asymptotically dominant term.

Theorem 5 (Instance-dependent 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. Assume that with probability $\ge 1-\frac{\delta}{2}$, $n_{h,s,a} \ge n\bar{d}_m$ for all h, s, a for some $\bar{d}_m > 0$. Then with probability at least $1-\delta$, we have:*

$$|\hat{v}^\pi - v^\pi| \le O(\sum_{h=1}^{H} \sum_{s,a} d_h^\pi(s,a) \sqrt{\frac{\mathrm{Var}_{s' \sim P_{h+1}(\cdot|s,a)}[V_{h+1}^\pi(s')] \log \frac{HSAn}{\delta}}{n_{h,s,a}}} + \frac{H^2 S \log \frac{HSAn}{\delta}}{n\bar{d}_m})$$

The above translates into the following worst-case bound.

Corollary 6 (Worst-case 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. Assume that* *with probability* $\geq 1-\frac{\delta}{2}$, $n_{h,s,a} \geq n\bar{d}_m$ *for all* h, s, a *for some* $\bar{d}_m > 0$. *Then with probability at least* $1-\delta$, *we have:*

$$|\hat{v}^\pi - v^\pi| \leq O(\sqrt{\frac{H^3 \log HSAn/\delta}{n\bar{d}_m}} + \frac{H^2 S \log \frac{HSAn}{\delta}}{n\bar{d}_m})$$

Inspired by Azar et al. (2017), Theorems 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 $(\hat{P}_{h+1} - P_{h+1})\hat{V}^\pi_{h+1}$. In order to treat the dominant term, we use Bernstein's inequality. To recover the worst-case bound in the corollary, we analyze the variance term with the canonical equality $\sum_h E_\pi[\mathrm{Var}_{s' \sim P(\cdot|s,a)}[V^\pi_h(s')]] = \mathrm{Var}_\pi[\sum_h r_h] - \sum_h \mathbb{E}_\pi[\mathrm{Var}[\mathbb{E}[r_h + V^\pi_{h+1}(s')|s,a]]] = O(H^2)$.

7 Empirical Studies

Our theoretical results certify that the TMIS estimator achieves low error even with adaptively logged data. However, they leave open some important question regarding the behavior of TMIS estimation under adaptive data.

1. When can adaptive logging help us improve our performance on offline tasks?
2. Can adaptive logging degrade our performance on offline tasks?
3. How does the choice of adaptivity affect our estimates? Can certain combinations of target policy and adaptivity (e.g. a highly suboptimal target policy and optimistic exploration), lead to positive or negative bias in $\hat{v}^\pi$? Our absolute bounds on estimation error do not resolve such questions.

   In this section, we probe these questions empirically, considering question 1 in **??** and question 2 in **??**.

7.1

Conclusion

This paper represents early steps in generalizing results from the non-adaptive setting to the adaptive setting. Though minimax-optimal worst-case bounds have been recovered in this paper, the instance-dependent results from Theorems 3 and 6 fail to recover the correct minimax behavior. This is because we could not salvage the martingale structure leveraged across timesteps in (Yin et al., 2021). However, based on our simulations and our intuition that throwing away data should not help us learn, we believe this gap to be an artifact of our analysis. As future work, we intend to recover better bounds by more carefully analyzing the 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.

We believe that theoretically clarifying the extent to which OPE methods carry over to the AOPE setting is an important step towards making offline RL a more convincing candidate for real-world decision problems.

References

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

Yaqi Duan, Zeyu Jia, and Mengdi Wang. Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, pages 2701–2709. PMLR, 2020.

Nathan Kallus, Yuta Saito, and Masatoshi Uehara. Optimal off-policy evaluation from multiple logging policies, 2020. URL `https://arxiv.org/abs/2010.11002`.

Nathan Lambert, Markus Wulfmeier, William Whitney, Arunkumar Byravan, Michael Bloesch, Vibhavari Dasagi, Tim Hertweck, and Martin Riedmiller. The challenges of exploration for offline reinforcement learning, 2022. URL `https://arxiv.org/abs/2201.11861`.

Wenlong Mou, Martin J. Wainwright, and Peter L. Bartlett. Off-policy estimation of linear functionals: Non-asymptotic theory for semi-parametric efficiency, 2022. URL `https://arxiv.org/abs/2209.13075`.

Jaehyeok Shin, Aaditya Ramdas, and Alessandro Rinaldo. Are sample means in multi-armed bandits positively or negatively biased?, 2019. URL `https://arxiv.org/abs/1905.11397`.

Chenjun Xiao, Ilbin Lee, Bo Dai, Dale Schuurmans, and Csaba Szepesvari. The curse of passive data collection in batch reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 8413–8438. PMLR, 2022.

Tengyang Xie, Yifei Ma, and Yu-Xiang Wang. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. *Advances in Neural Information Processing Systems*, 32, 2019.

Ming Yin and Yu-Xiang Wang. Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3948–3958. PMLR, 2020.

Ming Yin, Yu Bai, and Yu-Xiang Wang. Near-optimal provable uniform convergence in offline policy evaluation for reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1567–1575. PMLR, 2021.

A Concentration inequalities

Lemma A.1 (Hoeffding's Inequality)**.** *Let $x_1, ..., x_n$ bounded random variables such that $E[x_i] = 0$ and $|x_i| \leq \xi_i$ with probability 1. Then for any $\epsilon > 0$ we have:*

$$\Pr[|\frac{1}{n}\sum_{i=1}^{n} x_i| > \epsilon] \leq \exp\{-2\epsilon^2 n^2 / \sum_{i=1}^{n} \xi_i^2\}$$

Lemma A.2 (Bernstein's Inequality)**.** *Let $x_1, ..., x_n$ be independent bounded random variables such that $E[x_i] = 0$ and $|x_i| \leq \xi_i$ with probability 1. Let $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}[x_i]$, then with probability $1 - \delta$ we have:*

$$\frac{1}{n}\sum_{i=1}^{n} x_i \leq \sqrt{\frac{\sigma^2 \log\frac{1}{\delta}}{n}} + \frac{2\xi}{3n}\log\frac{1}{\delta}$$

Lemma A.3 (d-dimensional Concentration)**. ?** *Let z be a discrete random variable taking values in $\{1, ...d\}$. Let q be the the associated probability simplex. Assume we have n i.i.d. samples $z_1, ...z_n$, and define $\hat{q}$ by $\hat{q}_j = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{z_i = j\}}$. Then for any $\epsilon > 0$*

$$\Pr[\|\hat{q} - q\|_1 \geq \sqrt{d}(\frac{1}{\sqrt{n}} + \epsilon)] \leq \exp\{-N\epsilon^2\}$$

B The "Tape" View of Adaptivity

The simulation lemma allows us to decompose the error $\hat{v^\pi} - v^\pi$ into error terms corresponding to each possible (h, s, a)-tuple, $E_{h,s,a}$. $E_{h,s,a}$ will be instantiated differently across this paper, but it will always be a (data-independent) function, f, of the observed transitions out of (h, s, a), $\{s_{h+1}^i\}_{i:s_h^i = s, a_h^i = a}$, such as:

$$E_{h,s,a} = \frac{1}{n_{h,s,a}} \sum_{i:s_h^i = s, a_h^i = a} f(s_{h+1}^i) - \mathbb{E}[f(s')|s' \sim P_{h+1}(\cdot|s, a)]$$

We want to apply concentration inequalities to each $E_{h,s,a}$. We need to be careful though, because $\{s_{h+1}^i\}_{i=1}^n$ are mutually dependent. The crucial point in bounding $E_{h,s,a}$ is to observe that for fixed $n_{h,s,a}$, $E_{h,s,a}$ is amenable to concentration. More formally, we write:

$$E_{h,s,a} = \sum_{j=1}^{n} \mathbb{1}_{\{n_{h,s,a} = j\}} \frac{1}{n_{h,s,a}} \sum_{i:s_h^i = s, a_h^i = a} f(s_{h+1}^i) - \mathbb{E}[f(s')|s' \sim P_{h+1}(\cdot|s, a)] =$$

$$\sum_{j=1}^{n} \mathbb{1}_{\{n_{h,s,a} = j\}} \underbrace{\frac{1}{j}\sum_{i=1}^{j} f(s_{h+1}^{(i)}) - \mathbb{E}[f(s')|s' \sim P_{h+1}(\cdot|s, a)]}_{x_j}$$

where (i) re-indexes into trajectories that visit (s_h, a_h). Now each $f(s_{h+1}^{(i)})$ is independently distributed according to $f(s')$ with $s' \sim P_h(\cdot|s, a)$.

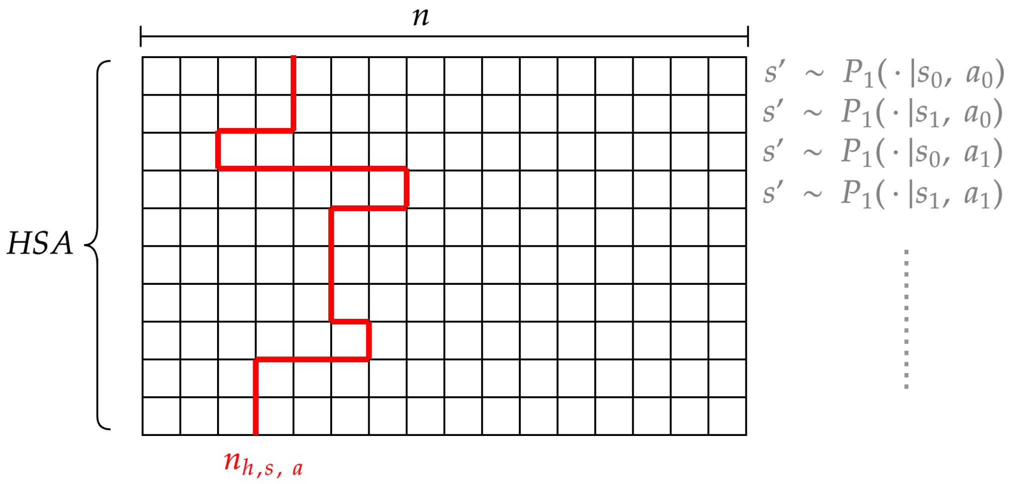
n
HSA
$s' \sim P_1(\cdot|s_0, a_0)$
$s' \sim P_1(\cdot|s_1, a_0)$
$s' \sim P_1(\cdot|s_0, a_1)$
$s' \sim P_1(\cdot|s_1, a_1)$
$n_{h,s,a}$

Figure 2: An illustration of the tape view of adaptivity. Each row (h, s, a) should be thought to contain n i.i.d. samples from $P_{h+1}(\cdot|s, a)$. The red "frontier" tracks how many samples have, in fact, been used by the logger (this quantity is always bounded by n).

In order to control $E_{h,s,a}$ with probabiliy δ, then, it suffices to control each x_j with probability δ/n. In this way, covering $n_{h,s,a}$ for all h, s, a yields bounds for the adaptive case. Note that in this paper the codomain of f will be either $\mathbb{R}$ or $\mathbb{R}^S$.

One way to visualize the dependence structure in $\mathcal{D}$ that we have leveraged above is to imagine a machine with a tape for each (h, s, a). The tape (h, s, a) contains an infinite string of i.i.d draws of $s' \sim P_h(\cdot|s, a)$, sampled before the logging begins. Every time (h, s, a) is visited by the logging algorithm, the s' read off of the corresponding tape, and the tape is advanced. Thus, a "frontier" is defined by $\{n_{h,s,a}\}_{h,s,a \in H \times S \times A}$ according to the logger's behavior. But, crucially, for any $n_{h,s,a}$, the transitions sampled off of the tape are i.i.d.

Whenever concentration is applied in this paper, the above covering argument is needed. Henceforth, every concentration argument will implicitly invoke the above covering trick, thus incurring a factor of n within the log term.

C Preliminaries and Proof of Uniform Bounds on $|\hat{v^\pi} - v^\pi|$

We assume for convenience that r and d_0 (the reward function and the initial distribution) are *known* in Appendices C and D. Because the error of $\hat{v}^\pi$ is dominated by error due to $\hat{P}$, this doesn't turn out to matter. For further discussion of this point see Appendix **??**

By the Simulation Lemma [cite], we have the following representation of the error, $|\hat{v^\pi} - v^\pi|$:

$$|\hat{v}^\pi - v^\pi| \leq \sum_{h=1}^{H-1} |\langle d_h^\pi, (\hat{P}_{h+1} - P_{h+1})\hat{V}_{h+1}^\pi \rangle| \leq \sum_{h=1}^{H-1} \sum_{s,a} |d_h^\pi(s,a)(\hat{P}_{h+1}(\cdot|s,a) - P_{h+1}(\cdot|s,a))^T \hat{V}_{h+1}^\pi| \tag{3}$$

where we remind the reader that $\hat{P}_t$ is the estimated transition kernel moving into the tth timestep, and $\hat{V}_t^\pi$ is the time-t value function for π under the approximate MDP defined by $\{\hat{P}_t\}_t$. Finally $d_t^\pi \in \mathbb{R}^{S\times A}$ is the marginal distribution of (s_t, a_t) under the true MDP and π. Theorem 3 follows quite directly from the above expansion, but we need to cover $n_{h,s,a}$ to take care of the dependence induced by the adaptivity.

In contrast to Yin and Wang (2020); Yin et al. (2021), our exploration assumption is levied on $\{n_{h,s,a}\}$ directly, rather than on the exploration parameters of the logging policy/ies (which would not sufficiently characterize the exploratory properties of an adaptive dataset). Thus, for each proof, we take care to budget for $\delta/2$ total failure probability across our concentration arguments – the remaining $\delta/2$ accounts for the failure case where $n_{h,s,a} = 0$ for some h, s, a.

C.1 Proof of Theorem 3 and Corollary 4

Using the third expression in equation 3, Hoelder's inequality, and concentration in 1-norm (Lemma A.3) we have (with probability at least $1 - \delta/2$):

$$|\hat{v}^\pi - v^\pi| \le \sum_{h=1} \sum_{s,a} d_h^\pi(s,a) \|\hat{P}_{h+1}(\cdot|s,a) - P_{h+1}(\cdot|s,a))\|_1 \|\hat{V}_{h+1}^\pi\|_\infty \le$$

$$\sum_{h=1}^{H-1} d_h^\pi(s,a) H (\sqrt{\frac{S\log\frac{HSAn}{\delta}}{n_{h,s,a}}} + \sqrt{\frac{S}{n_{h,s,a}}}) \le 2\sum_{h=1}^{H-1} d_h^\pi(s,a) H \sqrt{\frac{S\log\frac{HSAn}{\delta}}{n_{h,s,a}}}$$

where it must be noted that we have made the near-vacuous assumption that $n \ge e\delta \ge e\delta/HSA$ for which it suffices for n to be at least 3.

Note that all bounding was done independently of the policy, π. Therefore this result holds uniformly across all policies.

Corollary 4 follows immediately.

D Proof of Pointwise Bounds on $|\hat{v}^\pi - v^\pi|$

We now derive a slightly more sophisticated pointwise bound on the estimation error. We achieve this by replacing the estimated value function with the true value function in the simulation lemma, and pushing the residual into a lower-order term.

Again, we budget for $\delta/2$ failure probability, which allows us to provide for $n_{h,s,a} > 0$ for all h, s, a in what follows.

D.1 Proof of Theorem 5

Again, we start with Equation (3), which we re-express as

$$|\hat{v}^\pi - v^\pi| \le \sum_{h=1}^{H-1} \sum_{s,a} |d_h^\pi(s,a)[\underbrace{(\hat{P}_{h+1}(\cdot|s,a) - P_{h+1}(\cdot|s,a))^T V_{h+1}^\pi}_{*} + \underbrace{(\hat{P}_{h+1}(\cdot|s,a) - P_h(\cdot|s,a))^T(\hat{V}_{h+1}^\pi - V_{h+1}^\pi)}_{**}]$$

D.2 Bounding *

For arbitrary h, s, a, notice that

$$|(\hat{P}_{h+1}(\cdot|s,a) - P_{h+1}(\cdot|s,a))^T V^{\pi}_{h+1}| = |(\sum_{s'} (\hat{P}_{h+1}(s'|s,a) - P_{h+1}(s'|s,a))^T V^{\pi}_{h+1}(s')| =$$

$$|\sum_{s'} \frac{1}{n_{h,s,a}} (\sum_{i: s^i_h = s, a^i_h = a} \mathbb{1}_{\{s^i_{h+1} = s'\}} - P(s'|s,a)) V^{\pi}_{h+1}(s')|$$

$$\leq \frac{1}{n_{h,s,a}} \sum_{i: s^i_h = s, a^i_h = a} |V^{\pi}_{h+1}(s^i_{h+1}) - \mathbb{E}_{s^1_{h+1} \sim P_{h+1}(\cdot|s,a)}[V^{\pi}_{h+1}(s^1_{h+1})]|$$

As described in appendix B, we apply Bernstein's inequality while covering $n_{h,s,a}$ to get that with probability at least $1 - \delta/4HSAn$:

$$(\hat{P}_{h+1}(\cdot|s,a) - P_{h+1}(\cdot|s,a))^T V^{\pi}_{h+1} \leq \sqrt{\frac{2\mathrm{Var}[V^{\pi}_{h+1}(s')|s' \sim P_{h+1}(\cdot|s,a)] \log \frac{2HSAn}{\delta}}{n_{h,s,a}}} + \frac{4H}{3n_{h,s,a}} \log \frac{2HSAn}{\delta} \tag{4}$$

D.3 Bounding **

By Hoelder's inequality:

$$** \leq \|\hat{P}_{h+1}(\cdot|s,a) - P_{h+1}(\cdot|s,a)\|_1 \|\hat{V}^{\pi}_{h+1} - V^{\pi}_{h+1}\|_\infty$$

The first term above is bounded with the same S-dimensional concentration inequality that we used in C.1, Lemma A.3. With probability $1 - \delta/4HSA$, the following holds.

$$\|\hat{P}_{h+1}(\cdot|s,a) - P_{h+1}(\cdot|s,a)\|_1 \leq \sqrt{\frac{S \log \frac{4HSAn}{\delta}}{n_{h,s,a}}} + \sqrt{\frac{S}{n_{h,s,a}}} \leq 2\sqrt{\frac{S \log \frac{4HSAn}{\delta}}{n_{h,s,a}}} \leq 2\sqrt{\frac{S \log \frac{4HSAn}{\delta}}{n\bar{d}_m}} \tag{5}$$

where the final inequality reflects our exploration assumption on the logger. By the union bound, the above holds simultaneously for all h, s, a with probability at least $1 - \delta/4$.

We bound $|\hat{V}^{\pi}_{h+1} - V^{\pi}_{h+1}\|_\infty$ using the Simulation Lemma again. We needn't be too careful here, because $(**)$ will decay with $\frac{1}{n} \ll \frac{1}{\sqrt{n}}$, and so be dominated by $(*)$.

Fix h and s, and consider $\hat{V}^{\pi}_h(s) - V^{\pi}_h(s)$. Let $q^{(\pi,h,s)}_t(\cdot,\cdot)$ be the marginal visitation under policy π and MDP, whose dynamics are defined by P, but which starts deterministically from (h, s) for $t \geq h$.

$$\hat{V}^{\pi}_h(s) - V^{\pi}_h(s) = \sum_{t=h}^{H-1} \langle q^{(\pi,h,s)}_t(\cdot,\cdot), (\hat{P}_{t+1} - P_{t+1})\hat{V}^{\pi}_{t+1} \rangle \leq \sum_{t=h}^{H-1} \|q^{(\pi,h,s)}_t\|_1 \|\hat{P}_{t+1} - P_{t+1}\|_\infty \|\hat{V}^{\pi}_{h+1}\|_\infty$$

$$= \sum_{t=h}^{H-1} H \max_{s,a} \{\|(\hat{P}_{t+1} - P_{t+1})(\cdot|s,a)\|_1\} \leq 2H^2 \sqrt{\frac{S \log \frac{4HSAn}{\delta}}{n\bar{d}_m}}$$

where we have already provided for the control on $\max_{h,s,a}\{(\hat{P}_h(\cdot|s,a) - P_h(\cdot|s,a)\}$ in equation 5.

Combining the results of this section, we achieve

$$** \le \frac{4H^2 S \log\frac{HSAn}{\delta}}{n\bar{d}_m} \tag{6}$$

Combining the bounds on $(*)$ and $(**)$, we have that with probability at least $1-\delta/2$:

$$|\hat{v}^\pi - v^\pi| \le \sum_{h=1}^{H-1}\sum_{s,a} d_h^\pi(s,a)\left(\sqrt{\frac{2\mathrm{Var}[V_{h+1}^\pi(s')|s' \sim P_{h+1}(\cdot|s,a)]\log\frac{4HSAn}{\delta}}{n_{h,s,a}}} + \frac{4H}{3n\bar{d}_m}\log\frac{4HSAn}{\delta}\right) +$$

$$+ \frac{4H^3 S\log\frac{2HSAn}{\delta}}{n\bar{d}_m}$$

Whence

$$|\hat{v}^\pi - v^\pi| \le O(\sum_{h=1}^{H-1}\sum_{s,a} d_h^\pi(s,a)\sqrt{\frac{\mathrm{Var}[V_{h+1}^\pi(s')|s' \sim P_{h+1}(\cdot|s,a)]\log\frac{HSAn}{\delta}}{n_{h,s,a}}} + \frac{H^3 S\log\frac{HSAn}{\delta}}{n\bar{d}_m})$$

Observing that $\mathrm{Var}[V_{h+1}^\pi(s')] \le O(H^2)$ leads to corollary 6. For proof of this cute fact, see, for example, Lemma 3.4 in Yin and Wang (2020).

E On the sufficiency of considering d_0 and r to be known

In this section, we briefly justify our above assumption that $\hat{r} = r$ and $\hat{d}_0 = d_0$.

If, instead of using the true values d_0 and r, we use $\hat{d}_0$ and $\hat{r}$, the simulation lemma yields the following expansion of the error:

$$\hat{v}^\pi - v^\pi =$$

$$\langle \hat{d}_0 - d_0, \hat{V}_0^\pi\rangle + \sum_{h=1}^{H-1}\mathbb{E}_{s_h,a_h\sim d_h^\pi}[(\hat{P}_{h+1}(\cdot|s_h,a_h) - P_{h+1}(|s_h,a_h))^T\hat{V}_{h+1}^\pi] + \sum_{h=1}^{H-1}\mathbb{E}_{s_h,a_h\sim d_h^\pi}[\hat{r}_h(s_h,a_h) - r_h(s_h,a_h)]$$

Applying A.3 on the first term, we get a control on the form:

$$\langle \hat{d}_0 - d_0, \hat{V}_0^\pi\rangle \le \|\hat{d}_0 - d_0\|_1 \|\hat{V^\pi{}_h}\|_\infty \le H(\sqrt{S/n} + \sqrt{\frac{\log\frac{1}{\delta}}{n}}) \tag{7}$$

Furthermore, for all h,s,a, we can control the reward-term by applying Hoeffding's Inequality, Union Bound over (h,s,a), and covering of $n_{h,s,a}$:

$$\hat{r}_h(s,a) - r_h(s,a) = \frac{1}{n_{h,s,a}}\sum_{i:s_h^i=s,a_h^i=a} r_h^i - r_h(s,a) \le \sqrt{\frac{\log\frac{HSAn}{\delta}}{n_{h,s,a}}} \tag{8}$$

Both (7) and (8) can be absorbed into the bounds of Corollaries 3 and 5, at the cost of a constant factor.