Proceedings of Machine Learning Research vol vvv:1–23, 2025

Rates for Offline Reinforcement Learning with Adaptively Collected Data

Sunil Madhow SMADHOW@UCSD.EDU
Dan Qiao D2QIAO@UCSD.EDU
Ming Yin MY0049@PRINCETON.EDU
Yu-Xiang Wang YUXIANGW@UCSD.EDU

Abstract

Developing theoretical guarantees on the sample complexity of offline RL methods is an important step towards making data-hungry RL algorithms practically viable. Such results tend to 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 a relaxation of the i.i.d. setting that allows logging policies to depend adaptively upon previous data. For tabular MDPs, we show that minimax-optimal bounds on the sample complexity of offline policy evaluation (OPE) and offline policy learning (OPL) can be recovered under this adaptive setting, and also derive instance-dependent bounds. Finally, we conduct simulations to empirically analyze the behavior of these estimators under adaptive and non-adaptive data. We find that, even while controlling for logging policies, adaptive data can change the signed behavior of estimation error.

Keywords: RL, Offline RL, Off Policy Evaluation, Learning Theory

1. Introduction

Offline Reinforcement Learning (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 (Kiran et al., 2020) and medicine (Raghu et al., 2017), online RL algorithms are effectively ruled out by their dependence on randomized exploration. Even in lower-stakes applications, such as advertising (Cai et al., 2017), naively adopting online algorithms could mean throwing away vast reserves of previously-collected data. Efficient offline algorithms leave room for practitioners to exercise domain-specific control over the training process in a principled way.

Given a dataset, $\mathcal{D}$, of interactions with an MDP $\mathcal{M}$, two tasks that we may hope to achieve in offline RL are Offline Policy Evaluation (Yin and Wang, 2020) and Offline Learning (Lange et al., 2012). 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.

The theoretical question of how and when it is possible to perform OPE and OL given a specific dataset is the subject of much study (Lange et al., 2012; Raghu et al., 2018; Le et al.; Xie and Jiang, 2021; Duan et al., 2020; Yin and Wang, 2020; Yin et al., 2021; Jin et al., 2021; Yin et al., 2022; Qiao and Wang, 2023b; Zhang et al., 2022). Clearly, in order for $\mathcal{D}$ to be a rich enough dataset to learn from, strong assumptions need to be made about how well it explores the MDP. A standard assumption in offline RL is that $\mathcal{D}$ consists of i.i.d. trajectories distributed according

to some logging policy μ, where μ has "good" exploratory properties. However, it is difficult to justify the imposition of these assumptions on our data. How is a practitioner supposed to have run a "good" logging policy μ without a priori knowledge of the very MDP they aim to understand? In practice, the gathering of useful datasets is best done by running adaptive exploration algorithms (e.g. Lambert et al. (2022)), perhaps with human supervision and/or intervention. Any such scheme for data collection will necessarily be non-stationary and intradependent.

In this paper, we introduce the setting of Adaptive Offline RL (AORL), which allows datasets to be collected with arbitrary statistical drift in the logging policy that governs each trajectory. We show that existing techniques for policy evaluation and policy learning are efficient even in the AORL setting. In addition to the motivating examples given above, here are some scenarios that are covered by AORL but not by previous work:

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 (Schmucker et al., 2021).

2. The dataset $\mathcal{D}$ was gathered by humans, and therefore influenced by a number of unobserved factors and historical data. For example, a doctor prescribing medicine may make a determination based on her conversation with the patient and prior experience (Yu et al., 2019).

3. The dataset $\mathcal{D}$ has been gathered by a reward-free exploration algorithm (Jin et al., 2020; Wang et al., 2020; Qiao and Wang, 2023a). This dataset will have excellent exploratory properties, but is very intradependent.

1.1. Related Work

To the best of our knowledge, we are the first to consider OPE for reinforcement learning under adaptive data. However, in the study of bandits and RL, Off-Policy Evaluation has been an area of interest for more than a decade (Dudik et al., 2011; Jiang and Li, 2016; Wang et al., 2017; Thomas and Brunskill, 2016; Yin and Wang, 2020; Yin et al., 2021). In RL, existing work adopts the setting where $\mathcal{D} = \{\tau_i \sim \mu\}$ is a collection of i.i.d. trajectories. In this setting, bounds on the performance of OPE or OL algorithms are given in terms of an exploration parameter, like:

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

where $d_\cdot^\mu(\cdot,\cdot)$ is the marginal occupancy measure of μ. In the tabular, i.i.d. setting, OPE and OL may be considered solved problems (Le et al.; Duan et al., 2020; Yin et al., 2021; Xie and Jiang, 2021). The multi-logger setting, where $\mathcal{D} = \{\tau_i \sim \mu^i\}$ for μ^i statically chosen, is a straightforward generalization of the single-logger setting, and we refer to this problem as Non-Adaptive RL.

Especially relevant to our work is Yin et al. (2021), which derives the optimal rate for uniform OPE over the class of deterministic policies, and moreover establishes that uniform OPE in a neighborhood of an *empirically optimal* policy implies an optimal algorithm for the policy learning problem.

However, the practical value of bounds in these settings is still unclear. 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

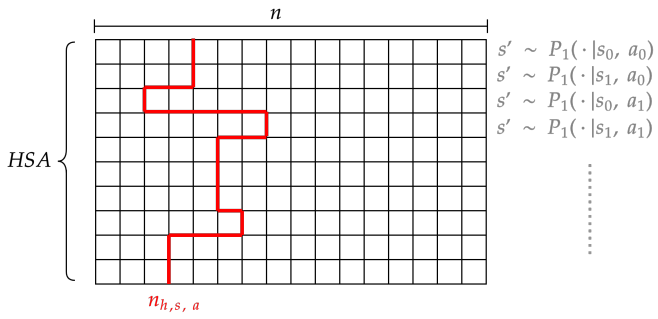

Figure 1: An illustration of the tape view of adaptive data collection. Each row (h, s, a) should be thought to contain n i.i.d. samples from $P_{h+1}(\cdot|s, a)$. We can view the logger as being provided with an adaptive subset of this table, defined by the red frontier above. The key point is that for fixed $n_{h,s,a}$, we may consider the transitions in the row (h, s, a) to be i.i.d.

knowledge of the MDP), they show a sample complexity exponential in H and S to be necessary in offline learning. Current results also fail to address the motivating application of learning from existing, human-generated data, which we would not expect to be identically distributed, or even independent (as the data collected in trajectory j almost certainly influences future policies $\mu^{j+1}, ...$).

While we do not know of any work that studies OPE with adaptively collected data, Jin et al. (2021) study the problem of Offline Learning with Pessimistic Value Iteration under adaptive data for linear MDPs. Their results do not cover OPE, and are loose when specialized to tabular MDPs. Jin et al. (2022) cover the problem of learning from adaptive data for contextual bandits. The generalization of such results to reinforcement learning is highly nontrivial, and the approach we take to the problem is largely unrelated. For multiarmed bandits, Shin et al. (2019) study how adaptive exploration schemes like optimism can lead to bias in estimated arm values. In their work, they imagine a data-collection model whereby a table is populated with data before any experiments begin. We make use of a similar model, generalized to the RL setting. This work also provides inspiration for our numerical simulations.

1.2. Novel Techniques and Contributions

We explain how to extend the results of Yin et al. (2021) to generate (near) minimax optimal solutions to uniform OPE and policy learning problems for the adaptive setting (Section 3). We derive *instance-dependent* bounds for uniform and pointwise OPE in the adaptive setting. Depending on the problem instance, these may give much faster rates than minimax bounds (Section 3). We empirically investigate the bias of model-based estimators under loggers that perform optimistic exploration (Section 4).

Our analysis introduces an equivalence between adaptive logging (Figure 2) and a "tape machine" model (Figure 1) which we believe has not been used in the RL setting. In particular, we view the logging process as adaptively querying entries of a table of transitions that has been prepopulated. The key point is that once we control the number of queries to each row, either by

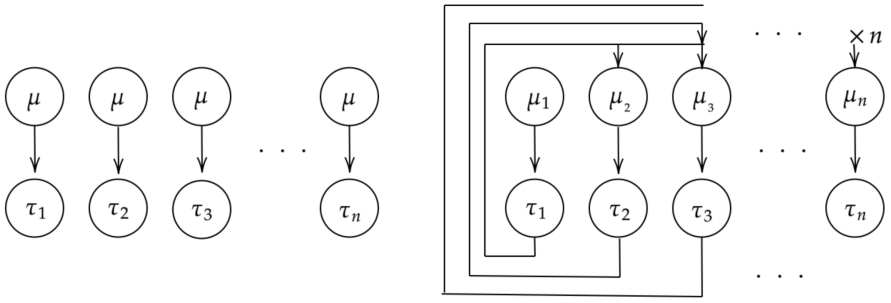

Figure 2: Non-adaptive regime (left) versus adaptive regime (right), depicted as a graphical model. We see that, in the adaptive regime, each policy depends on all previous trajectories. This induces dependence between the trajectories.

conditioning on them or covering all possibilities, we may treat the transitions within a row as iid (Appendix B).

2. Preliminaries

2.1. Symbols, notation, and MDP basics.

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

A Tabular, Finite-Horizon Markov Decision Process (MDP) is a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, r, P, d_1, H)$, where $\mathcal{S}$ is the discrete state space with $S := |\mathcal{S}|$, while $\mathcal{A}$ is the discrete action space with $A := |\mathcal{A}|$. Its dynamics are governed by a non-stationary 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 for all h, s, a.

$v^\pi := \mathbb{E}_\pi[\sum_{i=1}^H r_i | s_1 \sim d_1]$ is the value of the policy π, where the expectation is over the π-induced distribution over trajectories. Furthermore, we define for any π the value-function $V_h^\pi(s) := \mathbb{E}_\pi[\sum_{i=h}^H r_i | s_h = s]$ and Q-function $Q_h^\pi(s, a) := \mathbb{E}_\pi[\sum_{i=h}^H r_i | s_h = s, a_h = a]$ for $1 \le h \le H$.

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

2.2. Motivation and Problem Setup

Motivated both by the negative result from Xiao et al. (2022) and the complex structures of real-world data, 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 strong assumptions on d_m common to Duan et al. (2020); Yin et al. (2021), and the assumption of total ignorance found in Xiao et al. (2022). To this end, this paper studies the following problem:

Definition 1 (Adaptive Offline Reinforcement Learning) *An* adaptively collected *dataset $\mathcal{D}$ is a dataset of the form $\mathcal{D} = \{\tau_i \sim \mu^i\}_{i=1}^n$, where $\mu^1, ..., \mu^n$ are chosen adaptively. That is, μ^i may depend on the trajectories $\tau_1, ...\tau_{i-1}$ (Figure 2). The possibly random rule that chooses each μ^i is called an* adaptive logging algorithm.

Adaptive Offline RL *(AORL) refers to Offline RL where the dataset is assumed to have adaptively-collected. In particular* Adaptive Offline Policy Evaluation *(AOPE) and* Adaptive Offline Policy Learning *(AOPL) refer respectively to OPE and policy learning in the adaptive setting.*

As opposed to vanilla Offline RL, the AORL problem formulation allows for the data to have been collected according to a nearly arbitrary logging algorithm. 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, AORL addresses the key scenario of learning from intradependent, human-influenced datasets.

The issue of defining an exploration assumption for an adaptive logger is an interesting one. If $\mu^1, ...\mu^n$ were statically chosen,

$$\bar{d}_m := \frac{1}{n} \min_{h,s,a} \sum_{i=1}^{n} d_h^{\mu^i}(s,a) > 0 \tag{2}$$

would be a good assumption. However, the quantity $\bar{d}_m$ as defined above is now a random variable.

We find it most natural to levy our assumption directly on the number of visitations to each (h, s, a):

Assumption 2 (Exploration Assumption) *For $\bar{d}_m > 0$, logging process $\mathcal{E}$ satisfies a $(\bar{d}_m, \delta)$-exploration assumption if, with probability at least $1 - \delta$*

$$n_{h,s,a} > n\bar{d}_m$$

In the non-adaptive regime, $n_{h,s,a} \geq n\bar{d}_m/2$ holds for Equation 2's $\bar{d}_m$ by a multiplicative Chernoff bound. This implies that our results hold for the single-logger setting as a special case, ensuring this work is a strict generalization of single-logger theory. Furthermore, this assumption is generally satisfied by reward-free exploration algorithms (Jin et al., 2020; Qiao et al., 2022). Assumptions on the exploratory property of the logger will not always be necessary. Results that depend on Assumption 2 will state it as a hypothesis.

2.3. TMIS estimation

We consider the TMIS ("Tabular Marginalized Importance Sampling") 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

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

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 let:

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

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 of value function as

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

2.4. Burning data

For any dataset $\mathcal{D}$, we define $\mathcal{D}'$ to be the dataset that keeps only the first $N := \min_{h,s,a} n_{h,s,a}$ observations from each row of the tape machine. Let $\hat{w}^\pi$ be the TMIS estimator that is run on the burned dataset $\mathcal{D}'$. In light of Assumption 2, this transformation essentially reduces our setting to a generative model setting.

3. Theoretical Results

(Near)-Optimal Worst-case Bounds

We begin by observing that, as a consequence of the tape data model in Figure 1 (and Appendix B), we are able to recover (Yin et al., 2021)'s near-optimal worst-case bound for OPE and Offline learning.

Theorem 3 (Near optimal pointwise AOPE) *Suppose $\mathcal{D}$ is an adaptively-collected dataset (Definition 1) and Assumption 2 is satisfied with $(\bar{d}_m, \delta/2)$. Fix any policy π and let $\hat{w}^\pi$ be formed with the burned dataset $\mathcal{D}'$. Then with probability at least $1 - \delta$*

$$|\hat{v}^\pi - v^\pi| = \tilde{O}\left(\frac{H}{\sqrt{n\bar{d}_m}} + \frac{H^2\sqrt{SA}}{n\bar{d}_m}\right)$$

Proof Conditioned on $N := \min_{h,s,a} n_{h,s,a}$, we have that, for any $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$, $\{(s_{h+1}^{(i)}) | s_h^{(i)} = s, a_h^{(i)} = a, (i,h) \in \mathcal{D}'\}$ is a mutually independent set of draws from the distribution $P_h(\cdot|s,a)$. This is an immediate consequence of the tape-machine model (Appendix B).

Furthermore, by assumption, we have that $N > n\bar{d}_m$ with high probability. This observation implies that the Martingale-based proof in Appendix E of Yin et al. (2021), which conditions on N, is valid in the adaptive setting. ■

A union bound over deterministic policies above gives the optimal rate $\tilde{O}(\sqrt{\frac{H^3 S}{n\bar{d}_m}})$.

Using the same observation of conditional independence given N, we are able to reproduce the following result for policy learning.

Theorem 4 (Yin et al. (2021)-type near optimal offline learning result) *Suppose $\mathcal{D}$ is an adaptively-collected dataset satisfying Assumption 2 with parameters $(\bar{d}_m, \delta/2)$. For any policy ν, let $\hat{W}^\nu(\cdot)$ be its value function in the empirical MDP defined by the burned dataset, $\mathcal{D}'$. Let π^* be an optimal deterministic policy, $\hat{\pi}^* := \text{argmax}_\pi \hat{w}^\pi$ and $\hat{\pi}$ be such that $\|\hat{W}_t^{\hat{\pi}^*} - \hat{W}_t^{\hat{\pi}}\|_\infty \leq \xi$ for some $0 < \xi \leq \sqrt{H}/S$ and for all t. Then, with probability at least $1-\delta$, we have*

$$v^{\pi^*} - v^{\hat{\pi}} = \tilde{O}(\sqrt{\frac{H^3}{\bar{d}_m n}} + \xi)$$

Observe that we may generate $\hat{\pi}$ that satisfies Theorem 4's assumptions by running a planning algorithm like value iteration or policy iteration on the empirical MDP. Thu from a worst-case perspective, offline policy learning is solved for the adaptive setting using the same proof techniques as (Yin et al., 2021).

Instance-Dependent Upper Bounds

The minimax-optimal results from the previous section rely on the somewhat artificial trick of burning data to generate $\mathcal{D}'$. In this section, we provide instance-dependent upper bounds that do not require this trick, and may be much tighter for certain problem instances.

Theorem 5 (High-probability uniform bound on estimation error in AOPE) *Suppose $\mathcal{D}$ is an adaptively-collected dataset, and $\hat{v}^\pi$ is formed using this dataset. Then, with probability at least $1-\delta$, the following holds for all deterministic policies π:*

$$|\hat{v}^\pi - v^\pi| \leq \widetilde{O}\left(\sum_{h=1}^{H}\sum_{s,a} H d_h^\pi(s,a)\sqrt{\frac{S}{n_{h,s,a}}}\right),$$

where $n_{h,s,a}$ is the number of occurrences of (s_h, a_h) in $\mathcal{D}$ and with the convention that $\frac{0}{0} = 0$.

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 $\sqrt{H}$. Note that this result (and those that follow) use a $(\bar{d}_m, \delta/2)$-exploration assumption (Assumption 2).

Corollary 6 (High-probability uniform bound on estimation error in AOPE) *Suppose that $\mathcal{D}$, $\hat{v}^\pi$ are as in Theorem 5. Further suppose that Assumption 2 is satisfied with parameters $(\bar{d}_m, \delta/2)$. Then with probability $1-\delta$, we have that*

$$\sup_\pi |\hat{v}^\pi - v^\pi| \leq \widetilde{O}\left(H^2\sqrt{\frac{S}{n\bar{d}_m}}\right).$$

We also give a high-probability, instance-dependent, *pointwise* bound. In the pointwise case, we are able to shave off a $\sqrt{S}$ in the asymptotically dominant term.

Theorem 7 (Instance-dependent pointwise bound on estimation error in AOPE) *Fix a policy π. Suppose $\mathcal{D}$ is an adaptively collected dataset, and $\hat{v}^\pi$ is formed using this dataset. Further suppose that Assumption 2 is satisfied with parameters $(\bar{d}_m, \delta/2)$. Then with probability at least $1-\delta$, we have:*

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

The above translates into the following worst-case bound, which is suboptimal by a factor of $\sqrt{H}$.

Corollary 8 (Worst-case pointwise bound on estimation error in AOPE) *Consider the same setting as Theorem 7. Then with probability at least $1-\delta$, we have:*

$$|\hat{v}^\pi - v^\pi| \leq \widetilde{O}\left(\sqrt{\frac{H^3}{n\bar{d}_m}} + \frac{H^3 S}{n\bar{d}_m}\right).$$

The form of Theorems 5 and 7 reveal key features of a problem instance by decomposing estimation error along $[H]\times\mathcal{S}\times\mathcal{A}$. For example, $\mathrm{Var}_{s'\sim P_{h+1}(\cdot|s,a)}[V_{h+1}^\pi(s')]$ is a measure of how "pivotal" a transition is – that is, our uncertainty in the effect of the outcome of (h,s,a) on our future reward. $d_h^\pi(s,a)$ is, of course, a measure of how likely the target policy is to visit (h,s,a). Thus, these results tell us that it is important to design loggers that prioritize data-collection in regions of the MDP that are highly visited by π or highly pivotal.

Though Corollary 8 does not recover a minimax-optimal bound on the estimation error, $e^* := \widetilde{O}\left(\sqrt{\frac{H^2}{nd_m}}\right)$, Theorem 7 may be much tighter than e^* for certain MDPs or certain policies. To take an extreme case, Theorem 7 shows that error decreases as $O(\frac{1}{n})$ when the MDP is deterministic (a relevant setting in the alignment of diffusion models (Uehara et al., 2024)).

3.1. Proof sketches

Due to space constraints, we give only high-level sketches of the proofs in this section. Full proofs can be found in the appendices of the Arxiv version of this paper: https://arxiv.org/abs/2306.14063

High-probability Results: 5, 6, 7, 8

All of these results arise from applying concentration inequalities to functions of our estimates, $\hat{P}, \hat{r}, \hat{d}_0$. However, the first two quantities are formed using a mutually dependent dataset. The martingale structure of $|\hat{v}^\pi - v^\pi|$ used in Yin et al. (2021) is also lost in the adaptive setting, so there is no straightforward way to apply concentration. However, the tape model tells us that for fixed

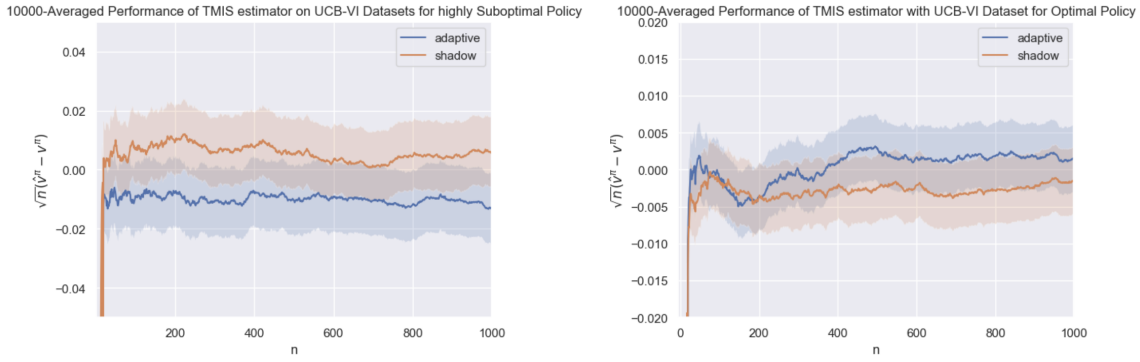

Figure 3: For different π, the blue curves show the average value of $\sqrt{n} \times (\hat{v}^\pi - v^\pi)$, where $\hat{v}^\pi$ is computed on the first $n \leq N$ trajectories of each of the $10,000$ adaptive datasets. The orange curves are computed in the same way, except using the $10,000$ shadow datasets. On the lefthand side, π is very suboptimal. On the righthand side, π is optimal. Confidence intervals are 95% Gaussian.

h, s, a and $n_{h,s,a}$, the error admits an expression amenable to concentration. Therefore, by using a covering argument over $\{n_{h,s,a}\}$, we are able to obtain our bounds while paying $HSAn$ inside the logarithm. Appendix B covers these details more carefully. Intuitively, the key observation is that even though the logging algorithm can query points $P_h(\cdot|s_h, a_h)$ adaptively, the transitions we observe are conditionally independent *once we control for the number of samples*.

The proof of Theorem 5 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 $\hat{d}_1$. The full proof is deferred to Appendix C.1.

Inspired by Azar et al. (2017), Theorem 7 is proved by applying concentration inequalities (with the same covering trick as Theorem 5) 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. The residual term scales with $\frac{1}{n} \ll \frac{1}{\sqrt{n}}$, which allows us to use cruder bounds when treating it. To recover the worst-case bound in the corollary, we analyze the variance term with the canonical equality $\sum_h E_\pi[\mathrm{Var}_{s'\sim P_h(\cdot|s,a)}[V^\pi_h(s')]] \leq \mathrm{Var}_\pi[\sum_h r_h] \leq H^2$. The full proof is presented in Appendix D.

4. Numerical Experiments

4.1. Experimental Motivation and Design

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

1. In multi-arm bandit literature, it has been established (Shin et al., 2019) that optimistic exploration causes negative bias in sample means for suboptimal arms. This motivates us to ask:

do datasets that are optimistically gathered lead to undervaluing sub-optimal policies in more complex MDPs?

2. Our results hold for arbitrary adaptivity, which may be adversarially chosen. But is it possible that some forms of adaptivity are beneficial? For example, suppose an optimistic logger collects a dataset $\mathcal{D}_a$ adaptively, with logging policies $\mu_1, ..., \mu_n$. Notice that $\mathcal{D}_a$ has a different distribution than a dataset $\mathcal{D}_b$ that consists of independent rollouts of $\mu_1, ..., \mu_n$. Is $\mathcal{D}_a$ more favorable for estimating high-value policies because it was adaptively guided towards high-value states?

We now investigate these questions empirically while validating our theoretical results. As our adaptive logger, we use UCB-VI. We gather data using UCB-VI, and then roll out an independent "Shadow" dataset using the same policies (Appendix G for details). As an optimistic algorithm, UCB-VI is well-suited to testing Question 1. Furthermore, as UCB-VI steers the data-collection procedure towards high-value states, it is conceivable that our estimator will benefit from the adaptivity for optimal π. UCB-VI can "react" to unwanted outcomes in trajectory τ_i, where the Shadow dataset cannot.

4.2. Results

We first consider a highly sub-optimal target policy. The lefthand side of Figure 3 shows two curves. Each curve plots the $\sqrt{n}$-scaled estimation error $\sqrt{n}(\hat{v}^\pi - v^\pi)$ against n, averaged over $10,000$ runs of the data-collection process ($10,000$ runs of UCB-VI and each run's corresponding shadow dataset). For each n and for each curve, this average is computed with respect to the first $n \leq N$ trajectories in each dataset. Theorem 6 tells us that these quantities will live in a band around zero, but does not give us information on their sign. Plotting the 95% confidence interval around each curve, there appears to be a distinction between the signed behaviors of the estimator for adaptive data vs non-adaptive data, though both confidence intervals cover 0. This suggests that there are measurable differences in the signed behavior of $\hat{v}^\pi$ when adaptive data is used, even if the logging policies are the same. Figure 3 righthand side suggests this does not happen for high-value policies.

On the whole, it seems that UCB-VI leads to negative bias in our estimates (especially for suboptimal policies), but limitations on the computational resources available to us constrain us to showing weak evidence of this conjecture. We also note that the magnitude of the error is indistinguishable for adaptive and non-adaptive data and that these simulations act as empirical validations of the results in Section 3, by showing that $\sqrt{n}|\hat{v}^\pi - v^\pi|$ does not explode.

5. Conclusion

In this paper, we derive upper bounds on the sample complexity of offline policy evaluation and offline policy learning, where the data-generating process can drift adaptively over time. This is facilitated by an argument from the "tape machine" model for data collection, which allows us to port results from the iid setting into the adaptive setting. In order to understand the dependence of estimation error on problem-specific quantities, we derive instance-dependent upper bounds and conduct an empirical simulations. In this paper, we only treated tabular MDPs; extending these methodologies to MDPs with continuous state spaces is left as important future work.

References

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

Han Cai, Kan Ren, Weinan Zhang, Kleanthis Malialis, Jun Wang, Yong Yu, and Defeng Guo. Real-time bidding by reinforcement learning in display advertising. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, 2017.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 30, 2017.

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

Miroslav Dudik, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *International Conference on Machine Learning*, 2011.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, 2016.

Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, 2020.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, 2021.

Ying Jin, Zhimei Ren, Zhuoran Yang, and Zhaoran Wang. Policy learning "without" overlap: Pessimism and generalized empirical bernstein's inequality. *Annals of Statistics*, 2022.

B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A. Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep reinforcement learning for autonomous driving: A survey, 2020. URL https://arxiv.org/abs/2002.00444.

Nathan Lambert, Markus Wulfmeier, William F. Whitney, Arunkumar Byravan, Michael Bloesch, Vibhavari Dasagi, Tim Hertweck, and Martin A. Riedmiller. The challenges of exploration for offline reinforcement learning. *CoRR*, abs/2201.11861, 2022.

Sascha Lange, Thomas Gabel, and Martin Riedmiller. Batch reinforcement learning. In *Reinforcement learning*, pages 45–73. Springer, 2012.

Hoang Minh Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *International Conference on Machine Learning*.

Dan Qiao and Yu-Xiang Wang. Near-optimal deployment efficiency in reward-free reinforcement learning with linear function approximation. In *International Conference on Learning Representations*, 2023a.

Dan Qiao and Yu-Xiang Wang. Offline reinforcement learning with differential privacy. In *Advances in Neural Information Processing Systems*, 2023b.

Dan Qiao, Ming Yin, Ming Min, and Yu-Xiang Wang. Sample-efficient reinforcement learning with loglog(T) switching cost. In *International Conference on Machine Learning*, pages 18031–18061. PMLR, 2022.

Aniruddh Raghu, Matthieu Komorowski, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach. In *Machine Learning for Healthcare Conference*, pages 147–163, 2017.

Aniruddh Raghu, Omer Gottesman, Yao Liu, Matthieu Komorowski, Aldo Faisal, Finale Doshi-Velez, and Emma Brunskill. Behaviour policy estimation in off-policy policy evaluation: Calibration matters. 2018.

Robin Schmucker, Jingbo Wang, Shijia Hu, and Tom M. Mitchell. Assessing the performance of online students – new data, new approaches, improved accuracy, 2021.

Jaehyeok Shin, Aaditya Ramdas, and Alessandro Rinaldo. Are sample means in multi-armed bandits positively or negatively biased? In *Advances in Neural Information Processing Systems*, 2019.

Philip S. Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning, 2016.

Masatoshi Uehara, Yulai Zhao, Tommaso Biancalani, and Sergey Levine. Understanding reinforcement learning-based fine-tuning of diffusion models: A tutorial and review, 2024. URL https://arxiv.org/abs/2407.13734.

Ruosong Wang, Simon S Du, Lin Yang, and Russ R Salakhutdinov. On reward-free reinforcement learning with linear function approximation. *Advances in neural information processing systems*, 33:17816–17826, 2020.

Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudik. Optimal and adaptive off-policy evaluation in contextual bandits, 2017.

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*, 2022.

Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. In Marina Meila and Tong Zhang, editors, *International Conference on Machine Learning*, 2021.

Ming Yin and Yu-Xiang Wang. Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, 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*, 2021.

Ming Yin, Yaqi Duan, Mengdi Wang, and Yu-Xiang Wang. Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism, 2022.

Chao Yu, Jiming Liu, and Shamim Nemati. Reinforcement learning in healthcare: A survey, 2019. URL https://arxiv.org/abs/1908.08796.

Ruiqi Zhang, Xuezhou Zhang, Chengzhuo Ni, and Mengdi Wang. Off-policy fitted q-evaluation with differentiable function approximators: Z-estimation and inference theory. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.