On Instance-Dependent Bounds for Offline Reinforcement Learning with Linear Function Approximation

Thanh Nguyen-Tang[1]*, **Ming Yin**[2], **Sunil Gupta**[3], **Svetha Venkatesh**[3], **Raman Arora**[1]

[1]Department of Computer Science, Johns Hopkins University
[2]Department of Computer Science and Department of Statistics and Applied Probability, UC Santa Barbara
[3]Applied AI Institute, Deakin University

Abstract

Sample-efficient offline reinforcement learning (RL) with linear function approximation has been studied extensively recently. Much of the prior work has yielded instance-independent rates that hold even for the worst-case realization of problem instances. This work seeks to understand instance-dependent bounds for offline RL with linear function approximation. We present an algorithm called Bootstrapped and Constrained Pessimistic Value Iteration (BCP-VI), which leverages data bootstrapping and constrained optimization on top of pessimism. We show that under a partial data coverage assumption, that of concentrability with respect to an optimal policy, the proposed algorithm yields a fast rate for offline RL when there is a positive gap in the optimal Q-value functions, even if the offline data were collected adaptively. Moreover, when the linear features of the optimal actions in the states reachable by an optimal policy span those reachable by the behavior policy and the optimal actions are unique, offline RL achieves absolute zero sub-optimality error when the number of episodes exceeds a (finite) instance-dependent threshold. To the best of our knowledge, these are the first results that give a fast rate bound on the sub-optimality and an absolute zero sub-optimality bound for offline RL with linear function approximation from adaptive data with partial coverage. We also provide instance-agnostic and instance-dependent information-theoretical lower bounds to complement our upper bounds.

Introduction

We consider the problem of offline reinforcement learning (offline RL), where the goal is to learn an optimal policy from a fixed dataset generated by some unknown behavior policy (Lange, Gabel, and Riedmiller 2012; Levine et al. 2020). The offline RL problem has recently attracted much attention from the research community. It provides a practical setting where logged datasets are abundant but exploring the environment can be costly due to computational, economic, or ethical reasons. It finds applications in a number of important domains including healthcare (Gottesman et al. 2019; Nie, Brunskill, and Wager 2021), recommendation systems (Strehl et al. 2010; Thomas et al. 2017; Zhang et al. 2022a), econometrics (Kitagawa and Tetenov 2018; Athey and Wager 2021), and more.

A large body of literature is devoted to providing generalization bounds for offline reinforcement learning with linear function approximation, wherein the reward and transition probability functions are parameterized as linear functions of a given feature mapping. For such linear MDPs, Jin, Yang, and Wang (2021) present a pessimistic value iteration (PEVI) algorithm and show that it is sample-efficient. In particular, Jin, Yang, and Wang (2021) provide a sample complexity bound for PEVI such that under the assumption that each trajectory is independently sampled and the behaviour policy is uniformly explorative in all dimensions of the feature mapping, the complexity bound improves to $\tilde{\mathcal{O}}(\frac{d^{3/2}H^2}{\sqrt{K}})$ where d is the dimension of the feature mapping, H is the episode length, and K is the number of episodes in the offline data. In a follow-up work, Xiong et al. (2023); Yin et al. (2022) leverage variance reduction (to derive a variance-aware bound) and data-splitting (to circumvent the uniform concentration argument) to further improve the result in Jin, Yang, and Wang (2021) by a factor of $\mathcal{O}(\sqrt{d}H)$. Xie et al. (2021) propose a pessimistic framework with general function approximation, and their bound improves that of (Jin, Yang, and Wang 2021) by a factor of $\sqrt{d}$ when the action space is finite, and the function approximation is linear. Uehara and Sun (2022) also obtain the $\frac{1}{\sqrt{K}}$ rate for offline RL with general function approximation, but like (Xie et al. 2021), their results are, in general, not computationally tractable as they require an optimization subroutine over a general function class. Although the $\frac{1}{\sqrt{K}}$ rate is minimax-optimal, in practice, assuming a worst-case setting is too pessimistic. Indeed, several empirical works suggest that in such natural settings, we can learn at a rate that is much faster than $\frac{1}{\sqrt{K}}$ (e.g., see Figure 1 in the supplementary). We argue that to circumvent these lower bounds and explain the rates we observe in practical settings, we should consider the intrinsic instance-dependent structure of the underlying MDP. Furthermore, most existing works establishing the the minimax-optimal $\frac{1}{\sqrt{K}}$ rate still require a strong assumption of uniform feature coverage and trajectory independence. This motivates us to study tighter instance-dependent bounds for offline RL with the mildest data cov-

*Email: nguyent@cs.jhu.edu / thnguyentang@gmail.com

erage condition possible.

Instance/gap-dependent bounds have been extensively studied in *online* bandit and reinforcement learning literature (Simchowitz and Jamieson 2019; Yang, Yang, and Du 2021; Xu, Ma, and Du 2021; He, Zhou, and Gu 2021). These works typically rely on an instance-dependent quantity, such as the minimum positive sub-optimality gap between an optimal action and the sub-optimal ones. However, to the best of our knowledge, it is still largely unclear how to leverage such an instance-dependent structure to improve offline RL, especially due to the unique challenge of distributional shift in offline RL as compared to the online case. A few recent works (Hu, Kallus, and Uehara 2021; Wang, Cui, and Du 2022) give gap-dependent bounds for offline RL; however, these works either require a strong uniform feature coverage assumption or only work for tabular MDPs. In addition, they require that the trajectories are collected independently across episodes – an assumption that is not very realistic as the data might have been collected by some online learning algorithms that interact with the MDPs (Fu et al. 2020). We are unaware of any existing work that leverages an instance/gap-dependent structure for offline RL with adaptive data and linear function approximation, which motivates the following question we consider in this paper.

Can we derive instance/gap-dependent bounds for offline RL with linear representations?

We answer the above question affirmatively and thus narrow the literature gap that were discussed in the recent work of (Wang, Cui, and Du 2022). In particular, we use $\Delta_{\min}$ to denote the minimum positive sub-optimality gap between the optimal action and the sub-optimal ones (Simchowitz and Jamieson 2019; Yang, Yang, and Du 2021; He, Zhou, and Gu 2021). The larger the $\Delta_{\min}$, the faster we can learn in an online setting since the actions with larger rewards are likely to be optimal, thereby reducing the time needed for exploration. Similarly, offline learning with uniform data coverage can benefit from the gap information as the entire state-action space is already fully explored by the offline policy (Hu, Kallus, and Uehara 2021). However, it remains elusive as *how an offline learner can benefit from the gap information where the learner cannot explore the environment anymore, and the offline data does not fully cover the state-action space*.

Our Contributions

We propose a novel bootstrapped and constrained pessimistic value iteration (BCP-VI) algorithm to leverage the gap information for an offline learner under partial data coverage, adaptive data, and linear function approximation. The key idea is to apply constrained optimization to the pessimistic value iteration (PEVI) algorithm of Jin, Yang, and Wang (2021) to ensure that each policy estimate has the same support as the behaviour policy. We then repeatedly apply the resulting algorithm to a sequence of partial splits bootstrapped from the original data to form an ensemble of policy estimates. Our key contributions are as follows.

1. We show that BCP-VI adapts to the instance-dependent quantity, $\Delta_{\min}$, to achieve a fast rate of $\mathcal{O}(\frac{\log K}{K})$, where K is the number of episodes in the offline data. Our result holds under the single-policy concentration coverage even when the offline data were collected adaptively.
2. As a byproduct, we also derive data-adaptive bounds for offline RL with linear function approximation under the single-policy concentrability assumption, which readily turns into a $\frac{1}{\sqrt{K}}$-bound with the single-policy concentration coefficients (without the gap information).
3. Under an additional condition that the linear features for optimal actions in states reachable by the behavior policy span those in states reachable by an optimal policy, we show that the policies returned by BCP-VI obtain a zero sub-optimality when K is larger than some problem-dependent constant.
4. We accompany our main result with information-theoretic lower bounds, which show that our gap-dependent bounds for offline RL are nearly optimal up to a polylog factor in terms of K and $\Delta_{\min}$. We summarize our results in Table 1.

Related Work

Offline RL with (linear) function approximation. While there has been much focus on provably efficient RL under linear function approximation, Jin, Yang, and Wang (2021) were the first to show that pessimistic value iteration is provably efficient for offline linear MDPs. Xiong et al. (2023) and Yin et al. (2022) improve upon Jin, Yang, and Wang (2021) by leveraging variance reduction and data splitting. Xie et al. (2021) consider a Bellman-consistency assumption with general function approximation, which improves the bound of Jin, Yang, and Wang (2021) by a factor of $\sqrt{d}$ when realized to finite action spaces and linear MDPs. On the other hand, Wang, Foster, and Kakade (2021) study the statistical hardness of offline RL with linear representation, suggesting that only realizability and strong uniform data coverage are insufficient for sample-efficient offline RL. Beyond linearity, the sample complexity of offline RL were studied with general, nonparametric or parametric, function approximation, typically based on Fitted-Q Iteration (FQI) (Munos and Szepesvári 2008; Le, Voloshin, and Yue 2019; Chen and Jiang 2019; Duan, Jin, and Li 2021; Duan, Wang, and Wainwright 2021; Hu, Kallus, and Uehara 2021; Hu et al. 2021; Nguyen-Tang et al. 2022b; Ji et al. 2023) or pessimism principle (Uehara and Sun 2022; Nguyen-Tang et al. 2022a; Jin, Yang, and Wang 2021; Xie et al. 2021; Nguyen-Tang and Arora 2023). However, all of the results above yield a worst-case bound of $\frac{1}{\sqrt{K}}$ without taking into account the structure of a problem instance.

Instance-dependent bounds for offline RL. The gap assumption (Assumption 3) has been studied extensively in online RL (Bubeck and Cesa-Bianchi 2012; Lattimore and Szepesvári 2020), yielding gap-dependent logarithmic regret bounds for bandits, tabular MDPs (Yang, Yang, and Du 2021) and MDPs with linear representation (He, Zhou, and Gu 2021). In online RL, when learning MDPs with linear rewards, under an additional assumption that the linear

Algorithm	Condition	Upper Bound	Lower Bound	Data
PEVI	Uniform	$\tilde{\mathcal{O}}\left(H^2 d^{3/2} K^{-1/2}\right)$	$\Omega\left(HK^{-1/2}\right)$	Independent
BCP-VI	OPC	$\tilde{\mathcal{O}}\left(H^2 d^{3/2} \kappa_* K^{-1/2}\right)$	$\Omega\left(H \kappa_{\min}^{1/2} K^{-1/2}\right)$	Adaptive
	OPC, Δ_{min}	$\tilde{\mathcal{O}}\left(d^3 H^5 \kappa_*^3 \Delta_{\min}^{-1} K^{-1}\right)$	$\Omega\left(H^2 \kappa_{\min} \Delta_{\min}^{-1} K^{-1}\right)$	Adaptive
	OPC, Δ_{min}, UO-SF, $K \geq k^*$	0	0	Adaptive
BCP-VTR	OPC	$\tilde{\mathcal{O}}\left(H^2 d \kappa_* K^{-1/2}\right)$	$\Omega\left(H \kappa_{\min}^{-1/2} K^{-1/2}\right)$	Adaptive
	OPC, Δ_{min}	$\tilde{\mathcal{O}}\left(d^2 H^5 \kappa_*^3 \Delta_{\min}^{-1} K^{-1}\right)$	$\Omega\left(H^2 \kappa_{\min} \Delta_{\min}^{-1} K^{-1}\right)$	Adaptive

Table 1: Bounds on the sub-optimality of offline RL with linear function approximation under different conditions and data coverage assumptions. The results in the first line were obtained in (Jin, Yang, and Wang 2021) under "sufficient" data coverage. Here, K is the number of episodes in the offline dataset, d is the dimension of the known linear mapping, H is the episode length, OPC stands for optimal policy concentrability (Assumption 1), $\kappa_* = \max_{h\in[H]} \kappa_h$ where κ_h is the OPC coefficient defined in Assumption 2, $\kappa_{\min} = \min_{h\in[H]} \kappa_h$, k^* is defined in Eq. (2), "Uniform" means uniform data coverage, "Independent" and "Adaptive" mean the episodes of the offline data were collected independently and adaptively, respectively, and UO-SF stands for unique optimality and spanning features in Assumption 4. BCP-VTR is a model-based offline RL method for linear mixture MDPs which is presented in the supplementary.

features of optimal actions span the space of the linear features of all actions (Papini et al. 2021), we can bound the regret by a constant. However, instance-dependent results for offline RL are still sparse and limited, mainly due to the unique challenge of distributional-shift in offline RL. There are only two instance-dependent works that we are aware of in the context of offline RL. The work of Hu, Kallus, and Uehara (2021) establishes a relationship between pointwise error rate of an estimate of Q^* and the rate of the resulting policy in Fitted Q-Iteration (FQI) and Bellman residual minimization under (a probabilistic version of) the minimum positive sub-optimality gap. Hu, Kallus, and Uehara (2021) showed that under the uniform feature coverage, i.e. $\lambda_{\min}\left(\mathbb{E}_{(s_h,a_h)\sim d_h^\mu}\left[\phi_h(s_h,a_h)\phi_h(s_h,a_h)^T\right]\right) > 0$ and the assumption that gap information is uniformly bounded away from zero with high probability, i.e. $\sup_\pi \mathbb{P}_{s\sim d^\pi}(0 < \Delta(s) < \delta) \leq (\delta/\delta_0)^\alpha$ for some constants $\delta_0 > 0, \alpha \in [0,\infty]$ and any $\delta > 0$, FQI yields a rate of $\mathcal{O}(\frac{1}{K})$ in linear MDP and $\mathcal{O}(e^{-K})$ in tabular MDP, respectively. A more recent work of Wang, Cui, and Du (2022) obtained gap-dependent bounds for offline RL; however, the results and technique (i.e. so-called the deficit thresholding technique) are limited only to independent data and tabular settings.

Offline RL from adaptive data. A common assumption for sample-efficient guarantees of offline RL is the assumption that the trajectories of different episodes are collected independently. However, it is quite common in practice that offline data is collected adaptively, for example, using contextual bandits, Q-learning, and optimistic value iteration. Thus, it is natural to study sample-efficient RL from adaptive data. Most initial results with adaptive data are for offline contextual bandits (Zhan et al. 2021a,b; Nguyen-Tang et al. 2022a; Zhang, Janson, and Murphy 2021). Pessimistic value iteration (PEVI) (Jin, Yang, and Wang 2021) works in linear MDP for the general data compliance assumption (see (Jin, Yang, and Wang 2021, Definition 2.1)), which is essentially equivalent to assuming that the data were adaptively collected. However, when deriving the explicit $\frac{1}{\sqrt{K}}$ bound of their algorithm, they made the assumption that the trajectories are independent (see their Corollary 4.6). The recent work of Wang, Cui, and Du (2022) derives a gap-dependent bound for offline tabular MDP but still requires that trajectories are collected independently.

Problem Setup

Episodic time-inhomogenous Markov decision processes (MDPs). A finite-horizon Markov decision process (MDP) is denoted as the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbb{P}, r, H, d_1)$, where $\mathcal{S}$ is an arbitrary state space, $\mathcal{A}$ is an arbitrary action space, H the episode length, and d_1 the initial state distribution. Let $\mathcal{P}(\mathcal{S})$ denote the set of probability measures over $\mathcal{S}$. A time-inhomogeneous transition kernel $\mathbb{P} = \{\mathbb{P}_h\}_{h=1}^H$, where $\mathbb{P}_h : \mathcal{S}\times\mathcal{A} \to \mathcal{P}(\mathcal{S})$ maps each state-action pair (s_h, a_h) to a probability distribution $\mathbb{P}_h(\cdot|s_h,a_h)$ (the corresponding density function $p_h(\cdot|s_h,a_h)$ is with respect to the Lebesgue measure ρ on $\mathcal{S}$). The reward function $r = \{r_h\}_{h=1}^H$, where $r_h : \mathcal{S}\times\mathcal{A} \to [0,1]$ is the mean reward function at step h. A policy $\pi = \{\pi_h\}_{h=1}^H$ assigns each state $s_h \in \mathcal{S}$ to a probability distribution, $\pi_h(\cdot|s_h)$, over the action space and induces a random trajectory $s_1, a_1, r_1, \ldots, s_H, a_H, r_H, s_{H+1}$ where $s_1 \sim d_1$, $a_h \sim \pi_h(\cdot|s_h)$, $s_{h+1} \sim \mathbb{P}_h(\cdot|s_h,a_h)$.

V-values and Q-values. For any policy π, the V-value function $V_h^\pi \in \mathbb{R}^{\mathcal{S}}$ and the Q-value function $Q_h^\pi \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ are defined as: $Q_h^\pi(s,a) = \mathbb{E}_\pi[\sum_{t=h}^H r_t | s_h = s, a_h = a]$, $V_h^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}[Q_h^\pi(s,a)]$. We also define $(P_hV)(s,a) := \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s,a)}[V(s')]$, $(T_hV)(s,a) := r_h(s,a) + (P_hV)(s,a)$. We have $Q_h^\pi = T_h V_{h+1}^\pi$ (the Bellman equation), $V_h^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}[Q_h^\pi(s,a)]$, $Q_h^* = T_h V_{h+1}^*$ (the Bellman optimality equation), and $V_h^*(s) =$

$\max_{a\in\mathcal{A}} Q_h^*(s,a)$. Let $\pi^* = \{\pi_h^*\}_{h\in[H]}$ be any deterministic, optimal policy, i.e., $\pi^* \in \arg\max_\pi Q^\pi$ and denote $v^* = v^{\pi^*}$. Moreover, let $d_h^{\mathcal{M},\pi}$ be the marginal state-visitation density for policy π at step h with respect to the Lebesgue measure ρ on $\mathcal{S}$, i.e., $\int_B d_h^{\mathcal{M},\pi}(s_h)\rho(ds_h) = \mathbb{P}(s_h \in B|d_1,\pi,\mathbb{P})$. We overload the notation $d_h^{\mathcal{M},\pi}(s_h,a_h) = d_h^{\mathcal{M},\pi}(s_h)\pi(a_h|s_h)$ for the state-action visitation density when the context is clear. We abbreviate $d_h^{\mathcal{M},*} = d_h^{\mathcal{M},\pi^*}$. Let $\mathcal{S}_h^{\mathcal{M},\pi} := \{s_h : d_h^{\mathcal{M},\pi}(s_h) > 0\}$ and $\mathcal{SA}_h^{\mathcal{M},\pi} := \{(s_h,a_h) : d_h^{\mathcal{M},\pi}(s_h,a_h) > 0\}$ be the set of feasible states and feasible state-action pairs, respectively, at step h under the policy π. Denote by $\mathcal{S}_h^{\mathcal{M}} = \cup_\pi \mathcal{S}_h^{\mathcal{M},\pi}$ and $\mathcal{SA}_h^{\mathcal{M}} = \cup_\pi \mathcal{SA}_h^{\mathcal{M},\pi}$ the set of all feasible states and feasible state-action pairs, respectively at step h. When the underlying MDP is clear, we drop the superscript $\mathcal{M}$ in $d^{\mathcal{M},\pi}$, $d^{\mathcal{M},*}$, $\mathcal{S}^{\mathcal{M},\pi}$, and $\mathcal{SA}^{\mathcal{M},\pi}$ and write d^π, d^*, $\mathcal{S}^\pi$, and $\mathcal{SA}^\pi$, respectively. We assume bounded marginal state(-action) visitation density functions and without loss of generality, we assume that $d_h^\pi(s_h,a_h) \le 1, \forall(h,s_h,a_h,\pi)$.[1]

Linear MDPs. When the state space is large or continuous, we often use a parametric representation for value functions or transition kernels. A standard parametric representation is linear models with given feature maps. In this paper, we consider such linear representation with the linear MDP (Yang and Wang 2019; Jin et al. 2020) where the transition kernel and the rewards are linear with respect to a given d-dimensional feature map: $\phi_h : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^d$.

Definition 1 (Linear MDPs)**.** *An MDP has a linear structure if for any (s,a,s',h),*

$$r_h(s,a) = \phi_h(s,a)^T\theta_h, \mathbb{P}_h(s'|s,a) = \phi_h(s,a)^T\mu_h(s'),$$

for some $\theta_h \in \mathbb{R}^d$ and some $\mu_h : \mathcal{S} \to \mathbb{R}^d$. For simplicity, we further assume that $\|\theta_h\|_2 \le \sqrt{d}$, $\|\int \mu_h(s)v(s)ds\|_2 \le \sqrt{d}\|v\|_\infty$ for any $v : \mathcal{S}\to\mathbb{R}$ and $\|\phi_h(s,a)\|_2 \le 1$.

Remark 1. *The linear MDP can be made practical with contrastive representation learning (Zhang et al. 2022b). We only consider linear MDP in the main paper but also consider a linear mixture model (Cai et al. 2020; Zhou, Gu, and Szepesvari 2021) in the supplementary.*

Offline Regime. In an offline learning setting, the goal is to learn a policy π that maximizes v^π given historical data, $\mathcal{D} = \{(s_h^t, a_h^t, r_h^t)\}_{h\in[H]}^{t\in[K]}$, generated by some unknown behaviour policy $\mu = \{\mu_h\}_{h\in[H]}$. Here, we allow the trajectory at any episode k to depend on the trajectories at all the previous episodes $t < k$. This reflects many practical scenarios where episode trajectories are collected adaptively by some initial online learner (e.g., ϵ-greedy, Q-learning, and LSVI-UCB).

In this paper, we assume that the support of $\mu_h(\cdot|s_h)$ for each s_h and h, denoted by $\text{supp}(\mu_h(\cdot|s_h))$, is known to the learner. We also denote the μ-supported policy class at stage h, denoted by $\Pi_h(\mu)$, as the set of policies whose supports belong to the support of the behavior policy:

$$\Pi_h(\mu) := \{\pi_h : \text{supp}(\pi_h(\cdot|s_h)) \subseteq \text{supp}(\mu_h(\cdot|s_h)), \forall s_h \in \mathcal{S}_h\}. \quad (1)$$

Performance metric. We define the sub-optimality of policy $\hat\pi$ as $\text{SubOpt}(\hat\pi) := \mathbb{E}_{s_1\sim d_1}[\text{SubOpt}(\hat\pi; s_1)]$, where $\text{SubOpt}(\hat\pi; s) := V_1^{\pi^*}(s_1) - V_1^{\hat\pi}(s_1)$. As $\hat\pi$ is learned from offline data $\mathcal{D}$, $\text{SubOpt}(\pi)$ is random (with respect to the randomness of $\mathcal{D}$ and possibly the internal randomness of the offline algorithm). The goal of offline RL is to learn $\hat\pi$ from $\mathcal{D}$ such that $\text{SubOpt}(\hat\pi)$ is small with high probability.

Bootstrapped and Constrained Pessimistic Value Iteration

We now describe the algorithm and establish instance-agnostic and instance-dependent bounds for offline RL from adaptive data with linear function approximation. With this algorithm, we show offline RL achieves a generic data-dependent bound under the optimal-policy concentrability assumption. Further, we adapt to the gap information giving an accelerated rate of suboptimality of $\frac{\log K}{K}$, and achieve zero sub-optimality when the optimal linear features under the behavior policy span those under an optimal policy.

Algorithm

We build upon the Pessimistic Value Iteration (PEVI) algorithm (Jin, Yang, and Wang 2021) with two essential modifications: bootstrapping and constrained optimization; hence the name Bootstrapped and Constrained Pessimistic Value Iteration (BCP-VI) in Algorithm 1. The constrained optimization on Line 10 ensures that the extracted policy is supported by the behaviour policy. The bootstrapping part divides the offline data in a progressively increasing split and applies the constrained version of PEVI in each split to form an ensemble (Line 14).[2]

Overall, BCP-VI estimates the optimal action-value functions Q_h^* leveraging its linear representation. In Line 6, it solves the regularized least-squares regression on $\mathcal{D}^{k-1}$:

$$\hat w_h := \arg\min_{w\in\mathbb{R}^d} \sum_{i=1}^{k} [\langle\phi(s_h^i,a_h^i), w\rangle - r_h^i - V_{h+1}(s_{h+1}^i)]^2 + \lambda\|w\|_2^2.$$

On Line 7, BCP-VI computes the action-value functions using $\hat w_h$, then offsets it with a bonus function b_h to ensure a pessimistic estimate. On Line 10, we extract $\hat\pi_h$ which is most greedy w.r.t. $\hat Q_h$ among the set of all policies $\Pi_h(\mu)$.

Policy execution. Given the policy ensemble $\{\hat\pi^k : k \in [K+1]\}$, we consider two ways of constructing the execution policy: creating a *mixture* $\hat\pi^{mix}$ and

[1] This trivially holds when $\mathcal{S}$ and $\mathcal{A}$ are discrete regardless of how large they are. When either $\mathcal{S}$ or $\mathcal{A}$ is continuous, we assume $d_h^\pi(s_h,a_h) \le B < \infty$ and assume $B = 1$ for simplicity.

[2] To be precise, this is not exactly bootstrapping in the traditional sense where the data is sampled with replacement and the ensemble is used to estimate uncertainty. Here we instead use progressive data splits to deal with adaptive data and form an ensemble of policy estimates.

Algorithm 1: Bootstrapped and Constrained Pessimistic Value Iteration (BCP-VI)

1: **Input:** Dataset $\mathcal{D} = \{(s_h^t, a_h^t, r_h^t)\}_{h\in[H]}^{t\in[K]}$, uncertainty parameters $\{\beta_k\}_{k\in[K]}$, regularization hyperparameter λ, μ-supported policy class $\{\Pi_h(\mu)\}_{h\in[H]}$.
2: **for** $k = 1, \ldots, K+1$ **do**
3: $\quad \hat{V}_{H+1}^k(\cdot) \leftarrow 0$.
4: $\quad$ **for** step $h = H, H-1, ..., 1$ **do**
5: $\quad\quad \Sigma_h^k \leftarrow \sum_{t=1}^{k-1} \phi_h(s_h^t, a_h^t) \cdot \phi_h(s_h^t, a_h^t)^T + \lambda \cdot I$.
6: $\quad\quad \hat{w}_h^k \leftarrow (\Sigma_h^k)^{-1} \sum_{t=1}^{k-1} \phi_h(s_h^t, a_h^t) \cdot (r_h^t + \hat{V}_{h+1}^k(s_{h+1}^t))$.
7: $\quad\quad b_h^k(\cdot,\cdot) \leftarrow \beta_k \cdot \|\phi_h(\cdot,\cdot)\|_{(\Sigma_h^k)^{-1}}$.
8: $\quad\quad \bar{Q}_h^k(\cdot,\cdot) \leftarrow \langle \phi_h(\cdot,\cdot), \hat{w}_h^k \rangle - b_h^k(\cdot,\cdot)$.
9: $\quad\quad \hat{Q}_h^k(\cdot,\cdot) \leftarrow \min\{\bar{Q}_h^k(\cdot,\cdot), H-h+1\}^+$.
10: $\quad\quad \hat{\pi}_h^k \leftarrow \arg\max_{\pi_h \in \Pi_h(\mu)} \langle \hat{Q}_h^k, \pi_h \rangle$
11: $\quad\quad \hat{V}_h^k(\cdot) \leftarrow \langle \hat{Q}_h^k(\cdot,\cdot), \pi_h^k(\cdot|\cdot) \rangle$.
12: $\quad$ **end for**
13: **end for**
14: **Output:** Ensemble $\{\hat{\pi}^k : k \in [K+1]\}$.

taking the policy $\hat{\pi}^{last}$ at the *last-iterate*; i.e, $\hat{\pi}^{mix} := \frac{1}{K}\sum_{k=1}^K \hat{\pi}^k$, and $\hat{\pi}^{last} := \hat{\pi}^{K+1}$. Note that $\hat{\pi}^{last}$ is similar to the PEVI policy in (Jin, Yang, and Wang 2021).

Practical considerations. In practice, when the action space is large, the constrained optimization on Line 10 can be relaxed to optimizing a regularized objective: $\max_{\pi_h} \langle \hat{Q}_h^k, \pi_h \rangle + \gamma \mathrm{KL}[\pi_h \| \mu_h]$ for some $\gamma > 0$. In settings where the behavior policy, μ, is not given, it can be simply estimated from the data. This relaxation assures that $\hat{\pi}_h^k$ is supported by μ_h and can be solved efficiently using an actor-critic framework. It is possible to include the optimization error of this actor-critic framework with a more involved analysis (Xie et al. 2021; Zanette, Wainwright, and Brunskill 2021; Cheng et al. 2022); we, however, ignore this here for simplicity.

Data-Dependent Bound

Sample-efficient offline reinforcement learning is not possible without certain data-coverage assumptions (Wang, Foster, and Kakade 2021). In this work, we rely on the optimal-policy concentrability (Assumption 1) which ensures that d^μ covers the trajectory of some optimal policy π^* and can be agnostic to other locations.

Assumption 1 (Optimal-Policy Concentrability (OPC) (Liu et al. 2020))**.** *There is an optimal policy π^*: $\forall (h, s_h, a_h)$, $d_h^{\pi^*}(s_h, a_h) > 0 \implies d_h^\mu(s_h, a_h) > 0$.*

Remark 2. *Consider any $s_h \in \mathcal{S}_h^{\pi^*}$. If $\pi_h^*(a_h|s_h) > 0$, then $d_h^{\pi^*}(s_h, a_h) > 0$, and thus $d_h^\mu(s_h, a_h) > 0$ by Assumption 1 which implies that $\mu_h(a_h|s_h) > 0$. For any $s_h \notin \mathcal{S}_h^{\pi^*}$, $\pi_h^*(\cdot|s_h)$ has no impact on the optimal value function $\{V_h^*\}_{h\in[H]}$. Thus, without loss of generality, we can assume that supp$(\pi_h^*(\cdot|s_h)) \subseteq$ supp$(\mu_h(\cdot|s_h)), \forall s_h \notin \mathcal{S}_h^{\pi^*}$. Overall, we have $\pi_h^* \in \Pi_h(\mu), \forall h \in [H]$.*

Assumption 1 is arguably the weakest data coverage assumption for sample-efficient offline RL, i.e., to ensure an optimal policy is statistically learnable from offline data (see the supplementary for a proof that the OPC condition is necessary). As such, Assumption 1 is significantly weaker than *uniform* data coverage assumption which features in most existing works in offline RL. The uniform feature coverage (Duan, Jia, and Wang 2020; Yin et al. 2022) requires that for all $h \in [H]$, $\lambda_{\min}\left(\mathbb{E}_{(s_h,a_h)\sim d_h^\mu}\left[\phi_h(s_h,a_h)\phi_h(s_h,a_h)^T\right]\right) > 0$, or $\min_{h,s_h,a_h} d_h^\mu(s_h, a_h) > 0$. The classical uniform concentrability (Szepesvári and Munos 2005; Chen and Jiang 2019; Nguyen-Tang et al. 2022b) requires that $\sup_{\pi,h,s_h,a_h} \frac{d_h^\pi(s_h,a_h)}{d_h^\mu(s_h,a_h)} < \infty$.

We further assume that the positive occupancy density under μ is bounded away from 0.

Assumption 2. $\kappa_h^{-1} := \inf_{(s_h,a_h):d_h^\mu(s_h,a_h)>0} d_h^\mu(s_h, a_h) > 0, \forall h \in [H]$.

Here, the infimum is over only the feasible state-action pairs under μ and it is agnostic to other locations. For example, the assumption is automatically satisfied when the state-action space is finite (albeit exponentially large, possibly). We remark that Assumption 2 is significantly milder than the uniform data coverage assumption $d_m := \inf_{h,s_h,a_h} d_h^\mu(s_h, a_h) > 0$ in (Yin, Bai, and Wang 2021b) as the infimum in the latter is uniformly over all states and actions.[3] Note that Assumption 2 also implies that $d_h^\mu(s_h) = \frac{d_h^\mu(s_h,a_h)}{\mu_h(a_h|s_h)} \geq \kappa_h^{-1}$ for any $s_h \in \mathcal{S}_h^\mu$. Combing with Assumption 1, κ_h can be seen as (an upper bound on) the *OPC coefficient* at stage h as we have $\frac{d_h^{\pi^*}(s_h,a_h)}{d_h^\mu(s_h,a_h)} \leq \kappa_h, \forall (h, s_h, a_h) \in [H] \times \mathcal{S} \times \mathcal{A}$.

Let $\delta \in (0, 1]$. Set $\lambda = 1$ for simplicity and $\beta_k = \beta_k(\delta) := c_1 \cdot dH \log(dHk/\delta)$ for some absolute constant $c_1 > 0, \forall k \in [K]$ in Algorithm 1. Then, we have the following data-dependent bound.

Theorem 1 (Data-dependent bound)**.** *Under Assumption 1, with probability at least $1 - 4\delta$ over the randomness of $\mathcal{D}$,*

$$
\begin{aligned}
&\mathrm{SubOpt}(\hat{\pi}^{mix}) \vee \mathrm{SubOpt}(\hat{\pi}^{last}) \\
&\leq \frac{4\beta(\delta)}{K} \sum_{h=1}^H \sum_{k=1}^K \frac{d_h^*(s_h^k, a_h^k)}{d_h^\mu(s_h^k, a_h^k)} \|\phi_h(s_h^k, a_h^k)\|_{(\Sigma_h^k)^{-1}} \\
&+ \frac{4\beta(\delta)}{K} \sum_{h=1}^H \sqrt{\log\left(\frac{H}{\delta}\right) \sum_{k=1}^K \left(\frac{d_h^*(s_h^k, a_h^k)}{d_h^\mu(s_h^k, a_h^k)}\right)^2} \\
&+ \frac{2}{K} + \frac{16H}{3K} \log\left(\frac{\log_2(KH)}{\delta}\right).
\end{aligned}
$$

[3] Interestingly, this assumption was also used, independently, in the prior work of Yin, Bai, and Wang (2021a).

Remark 3. *The first term in the bound in Theorem 1 is the elliptical potential that results from pessimism and is the dominant term, whereas the other terms are generalization errors resulting from the measure concentration phenomenon and the peeling technique.*

The sub-optimality bound in Theorem 1 explicitly depends on the observed data in the offline data via the marginalized density ratios $\frac{d_h^*(s_h^k, a_h^k)}{d_h^\mu(s_h^k, a_h^k)}$ (which is valid thanks to Assumption 1). One immediate consequence of the data-dependent bound in Theorem 1 is that the bound can turn into a weaker, yet, more explicit rate of $\frac{1}{\sqrt{K}}$.

Corollary 1. *Under Assumptions 1-2, with probability at least $1 - \Omega(1/K)$ over the randomness of $\mathcal{D}$, we have:*

$$\mathbb{E}\left[\mathrm{SubOpt}(\hat{\pi}^{mix})\right] \vee \mathbb{E}\left[\mathrm{SubOpt}(\hat{\pi}^{last})\right] = \tilde{\mathcal{O}}\left(\frac{\kappa H d^{3/2}}{\sqrt{K}}\right),$$

where $\kappa := \sum_{h=1}^H \kappa_h$.

Comparing with Yin and Wang (2021). Yin and Wang (2021) also use OPC to establish the intrinsic offline learning bound with pessimism and leverage the variance information to obtain a tight dependence on H. Their result is valid only for tabular MDPs with the finite state space and finite action space and cannot generalize to linear MDPs.

Comparing with Jin, Yang, and Wang (2021). Similarly, Jin, Yang, and Wang (2021) also consider linear MDPs with pessimism and provide a generic bound under arbitrary data coverage. They then realize their generic bound in the uniform feature coverage assumption (Duan, Jia, and Wang 2020; Yin and Wang 2021) to obtain a sub-optimality bound of $\tilde{\mathcal{O}}(\frac{d^{3/2}H^2}{\sqrt{K}})$. However, the uniform feature coverage is not necessary to obtain the $\frac{1}{\sqrt{K}}$ bound; in our result, we demonstrate that OPC is sufficient to get the $\frac{1}{\sqrt{K}}$ bound.

Comparing with Xie et al. (2021). Xie et al. (2021) consider Bellman-consistent pessimism for offline RL with general function approximation, where they maintain a version space of all functions that have small Bellman evaluation error and select a function from the version space that has the smallest initial value. Their algorithm is, however, computationally intractable in general. When realized to linear MDPs, they do have a tractable algorithm but its guarantee requires the behaviour policy to be explorative in all dimensions of the feature mapping, i.e., $\mathbb{E}_\mu[\phi(s,a)\phi(s,a)^T] \succ 0$. We do not require such an assumption in our analysis.

Theorem 1 is a byproduct that sets the stage for our instance-dependent bounds in the following section. Nonetheless, to the best of our knowledge, Theorem 1 is the first result to provide a $\frac{1}{\sqrt{K}}$ rate for linear MDPs with OPC.

Remark 4. *In the Appendix, we show that OPC is necessary to guarantee a sublinear sub-optimality bound for offline RL. When OPC fails to hold, we show in the appendix that BCP-VI suffers a constant sub-optimality incurred at optimal locations that are not supported by the behavior policy.*

Instance-Dependent Bounds

We now show that BCP-VI automatically exploits various types of instance-dependent structures of the underlying MDP to achieve an accelerated rate on the sub-optimality.

Gap-dependent bound. A natural measure of the hardness of an MDP instance is the minimum positive action gap (Assumption 3) which determines how hard it is to distinguish optimal actions from sub-optimal ones.

Definition 2. *For any $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, the sub-optimality gap $\Delta_h(s,a)$ is defined as: $\Delta_h(s,a) := V_h^*(s) - Q_h^*(s,a)$, and the minimal sub-optimality gap is defined as:*

$$\Delta_{\min} := \min_{s,a,h}\{\Delta_h(s,a) | \Delta_h(s,a) \neq 0\}.$$

We assume that the minimal sub-optimality gap is strictly positive, which is a common assumption for gap-dependent analysis (Simchowitz and Jamieson 2019; Yang, Yang, and Du 2021; He, Zhou, and Gu 2021).

Assumption 3. $\Delta_{\min} > 0$.

Theorem 2. *Under Assumptions 1-2-3, where $\kappa_* = \max_{h\in[H]} \kappa_h$, with probability at least $1 - (1 + 3\log_2(H/\Delta_{\min}))\delta$, we have that*

$$\mathrm{SubOpt}(\hat{\pi}^{mix}) \lesssim 2\frac{d^3H^5\kappa_*^3}{\Delta_{\min}\cdot K}\log^3(dKH/\delta) + \frac{16H\kappa_*}{3K}\log\log_2(KH\kappa_*/\delta) + \frac{2}{K}.$$

Remark 5. *If we set the δ in Theorem 2 as $\delta = \Omega(1/K)$, then for the expected sub-optimality bound, we have:*

$$\mathbb{E}\left[\mathrm{SubOpt}(\hat{\pi}^{mix})\right] = \tilde{\mathcal{O}}\left(\frac{d^3H^5\kappa_*^3}{\Delta_{\min}\cdot K}\right).$$

The bound in Theorem 2 depends on $\Delta_{\min}$ inversely. It is independent of the state space $\mathcal{S}$, action space $\mathcal{A}$, and is logarithmic in the number of episodes K. This suggests that our offline algorithm is sample-efficient for MDPs with large state and action spaces. This is the first result of its kind to leverage the gap information $\Delta_{\min}$ to obtain $\mathcal{O}\left(\frac{\log K}{K}\right)$ bound for offline RL with linear function approximation, partial data coverage and adaptive data.

We now provide the information-theoretic lower bound of learning offline linear MDPs under Assumptions 1-2-3.

Theorem 3. *Fix any $H \geq 2$. For any algorithm Algo($\mathcal{D}$), and any concentrability coefficients $\{\kappa_h\}_{h\geq 1}$ such that $\kappa_h \geq 2$, there exists a linear MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r, d_0)$ with a positive minimum sub-optimality gap $\Delta_{\min} > 0$ and dataset $\mathcal{D} = \{(s_h^t, a_h^t, r_h^t)\}_{h\in[H]}^{t\in[K]} \sim \mathcal{P}(\cdot|\mathcal{M}, \mu)$ where $\sup_{h,s_h,a_h} \frac{d_h^{\mathcal{M},*}(s_h,a_h)}{d_h^{\mathcal{M},\mu}(s_h,a_h)} \leq \kappa_h, \forall h \in [H]$ such that:*

$$\mathbb{E}_{\mathcal{D}\sim\mathcal{M}}\left[\mathrm{SubOpt}(\mathit{Algo}(\mathcal{D}); \mathcal{M})\right] = \Omega\left(\frac{\kappa_{\min}H^2}{K\Delta_{\min}}\right),$$

where $\kappa_{\min} = \min\{\kappa_h : h \in [H]\}$.

Theorem 3 implies that any offline algorithm suffers the expected sub-optimality of $\Omega\left(\frac{\kappa_{\min} H^2}{K\Delta_{\min}}\right)$ under a certain linear MDP instance and behaviour policy that satisfy the minimum positive action gap and the single concentrability. Thus, the result suggests that our algorithm is optimal in terms of K and $\Delta_{\min}$ up to log factors.

Zero sub-optimality. We introduce additional assumptions on the linear mapping which our algorithm can exploit to further accelerate the rate. Let $\text{span}(\mathcal{X})$ the vector space spanned by $\mathcal{X}$.

Assumption 4. *1. (Unique Optimality - UO): The optimal actions are unique, i.e.*

$$|\textit{supp}(\hat{\pi}_h^*(\cdot|s_h))| = 1, \forall (h, s_h) \in [H] \times \mathcal{S}_h^*.$$

2. (Spanning Features - SF): Let $\phi_h^*(s) := \phi_h(s, \pi_h^*(s))$. For any $h \in [H]$,

$$\textit{span}\{\phi_h^*(s_h) : \forall s_h \in \mathcal{S}_h^\mu\} \subseteq \textit{span}\{\phi_h^*(s_h) : \forall s_h \in \mathcal{S}_h^*\}.$$

Intuitively, the features of optimal actions in states reachable by an optimal policy provide all information about those in states reachable by the behaviour policy μ. Note that Assumption 4.2 is much milder than the uniform feature coverage assumption as it does not impose any constraint on the linear features with respect to the offline policy and does not require $\text{span}\{\phi_h^*(s) : \forall s \in \mathcal{S}, d_h^*(s) > 0\}$ to span the entire $\mathbb{R}^d$. In online regime, a similar assumption called "universally spanning optimal features" is used to obtain constant regrets (Papini et al. 2021). However, their assumption is strictly stronger than ours as they require $\text{span}\{\phi_h^*(s) : \forall s \in \mathcal{S}, d_h^*(s) > 0\}$ to span all the features of all actions and states reachable by *any* policy. Assumption 4.2 instead requires such condition only over optimal actions and states reachable by the behavior policy.

Let λ_h^+ be the smallest positive eigenvalue of $\Sigma_h^* := \mathbb{E}_{(s_h,a_h)\sim d_h^{\pi^*}}[\phi_h(s_h,a_h)\phi_h(s_h,a_h)^T]$, let $\kappa_{1:h} := \prod_{i=1}^h \kappa_i$, and define:

$$k^* = \max_{h\in[H]} \bar{k}_h \vee \tilde{k}_h, \tag{2}$$

where $\bar{k}_h := \tilde{\Omega}\left(\frac{d^6 H^4 \kappa^6}{\Delta_{\min}^4 (\lambda_h^+)^2} + \frac{\kappa_{1:h}}{\lambda_h^+}\right) \wedge \tilde{\Omega}\left(\frac{\kappa_{1:h}^2 \kappa^2 H^2 d^3}{(\lambda_h^+)^2}\right)$, $\tilde{k}_h := \tilde{\Omega}\left(\frac{d^2 H^4 \kappa_{1:h}}{\Delta_{\min}^2 (\lambda_h^+)^3}\right), \forall h.$

Theorem 4. *Given Assumptions 1-2-3-4, with probability at least $1-4\delta$, we have that $\text{SubOpt}(\hat{\pi}^k) = 0, \forall k \geq k^*$, where k^* is defined in Eq. (2).*

Remark 6. *The thresholding value k^* defined in Eq. (2) is independent of K, and it scales with the inverse of $\Delta_{\min}$ and the distributional shift measures κ_h.*

Theorem 4 suggests that when the linear feature at the optimal actions are sufficiently informative and when the number of episodes is sufficiently large exceeding an instance-dependent threshold specified by k^*, $\hat{\pi}^k$ precisely recovers the (unique) optimal policy with high probability.

Proof Overview

Here, we provide a brief overview of the key ideas from our proof technique; we defer the details to the Appendix.

Proof of Theorem 1. With the extended value difference and the constrained optimization in Line 10 of Algorithm 1, we reduce bounding $\text{SubOpt}(\hat{\pi}^k)$ to bounding $2\mathbb{E}_{\pi^*}[\sum_{h=1}^H b_h^k(s_h, a_h)]$. We then use the marginalized importance sampling to convert $2\mathbb{E}_{\pi^*}[\sum_{h=1}^H b_h^k(s_h, a_h)]$ to the dominant term $\beta(\delta)\sum_{h=1}^H \frac{d_h^*(s_h^k, a_h^k)}{d_h^\mu(s_h^k, a_h^k)} \|\phi_h(s_h^k, a_h^k)\|_{(\Sigma_h^k)^{-1}}$. For $\hat{\pi}^{last}$, the key observation is that $\Sigma_h^k \preceq \Sigma_h^{K+1}$, thus

$$2\mathbb{E}_{\pi^*}[\sum_{h=1}^H \|\phi_h(s_h, a_h)\|_{(\Sigma_h^{K+1})^{-1}}] \leq \frac{2}{K}\sum_{k=1}^K \mathbb{E}_{\pi^*}[\sum_{h=1}^H \|\phi_h(s_h, a_h)\|_{(\Sigma_h^k)^{-1}}].$$

Proof of Theorem 2. We relate bounding $\text{SubOpt}(\hat{\pi}^{mix})$ to bounding the empirical version $\frac{1}{K}\sum_{k=1}^K \text{SubOpt}(\hat{\pi}^k; s_1^k)$ plus an estimation error term. Using the original online-to-batch argument (Cesa-Bianchi, Conconi, and Gentile 2004) only gives a $\frac{1}{\sqrt{K}}$ generalization error which prevents us from obtaining $\tilde{\mathcal{O}}(\frac{1}{K})$ bound. Instead, we propose an improved online-to-batch argument (Lemma 5) with $\tilde{\mathcal{O}}(\frac{1}{K})$ generalization error; this may be of independent interest. Then, $\text{SubOpt}(\hat{\pi}^k; s_1)$ is expressed through decomposition $\text{SubOpt}(\hat{\pi}^k; s_1) = \mathbb{E}_{\hat{\pi}^k}[\sum_{h=1}^H \Delta_h(s_h, a_h)|\mathcal{F}_{k-1}, s_1]$ (Lemma 11). To handle the gap terms, the key observation is that $\hat{\pi}^k$ belongs to the μ-supported policy class (Lemma 12). Thus, the concentrability coefficients (Assumption 2) apply and so does the marginalized importance sampling. The next step is to count the number of times the empirical gaps exceed a certain value, $\sum_{k=1}^K \mathbf{1}\{\Delta_h(s_h^k, \hat{\pi}_h^k(s_h^k)) \geq \Delta\} \lesssim \frac{d^3 H^2 \iota^{-2}}{\Delta^2} \log^3(dKH/\delta)$ (Lemma 14).

Proof of Theorem 4. A key observation is that $\lambda_{\min}(\Sigma_h^k) \gtrsim k\lambda_h^+$ (Lemma 17) where λ_h^+ is the minimum positive eigenvalue of Σ_h^*. Thus, for any $v \in \text{span}(\{\phi_h^*(s)|s \in \mathcal{S}_h^*\})$, $\|v\|_{(\Sigma_h^k)^{-1}} \leq \mathcal{O}(1/\sqrt{k})$ (Lemma 18). Under Assumption 4, $\forall s_h \in \mathcal{S}_h^\mu$, $\Delta_h(s_h, \hat{\pi}_h^k(s_h)) \leq 2\beta_k \mathbb{E}_{\pi^*}[\sum_{h'=h}^H \|\phi_h(s_h, a_h)\|_{(\Sigma_{h'}^k)^{-1}}|\mathcal{F}_{k-1}, s_h]$
$= \mathcal{O}(\frac{1}{\sqrt{k}}) < \Delta_{\min}$, for sufficiently large k.

Proof of Theorem 3. We reduce the lower bound construction to statistical testing using the Le Cam method, and construct a hard MDP instance based on the construction of Jin, Yang, and Wang (2021) with a careful design of the behavior policy to incorporate the OPC coefficients and the gap information $\Delta_{\min}$.

Discussion

This work studies offline RL with linear function approximation and contributes a first-of-its-kind $\tilde{\mathcal{O}}(\frac{1}{K\Delta_{\min}})$ bound and a constant bound in this setting, using bootstrapping and constrained optimization on top of pessimism. A question that remains is to close the gap between upper bounds and lower bounds in terms of d and κ.

Acknowledgements

This research was supported, in part, by DARPA GARD award HR00112020004, NSF CAREER award IIS-1943251, an award from the Institute of Assured Autonomy, and Spring 2022 workshop on "Learning and Games" at the Simons Institute for the Theory of Computing. MY was partially supported by NSF Award #2007117 and #2003257. SV is the recipient of an ARC Australian Laureate Fellowship (FL170100006). We thank our anonymous reviewers at AAAI'23 for the constructive comments.

References

Athey, S.; and Wager, S. 2021. Policy learning with observational data. *Econometrica*, 89(1): 133–161.

Bubeck, S.; and Cesa-Bianchi, N. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*.

Cai, Q.; Yang, Z.; Jin, C.; and Wang, Z. 2020. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, 1283–1294. PMLR.

Cesa-Bianchi, N.; Conconi, A.; and Gentile, C. 2004. On the Generalization Ability of On-Line Learning Algorithms. *IEEE Trans. Inf. Theory*, 50(9): 2050–2057.

Chen, J.; and Jiang, N. 2019. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, 1042–1051. PMLR.

Cheng, C.-A.; Xie, T.; Jiang, N.; and Agarwal, A. 2022. Adversarially trained actor critic for offline reinforcement learning. In *International Conference on Machine Learning*, 3852–3878. PMLR.

Duan, Y.; Jia, Z.; and Wang, M. 2020. Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, 2701–2709. PMLR.

Duan, Y.; Jin, C.; and Li, Z. 2021. Risk bounds and rademacher complexity in batch reinforcement learning. In *International Conference on Machine Learning*, 2892–2902. PMLR.

Duan, Y.; Wang, M.; and Wainwright, M. J. 2021. Optimal policy evaluation using kernel-based temporal difference methods. *arXiv preprint arXiv:2109.12002*.

Fu, J.; Kumar, A.; Nachum, O.; Tucker, G.; and Levine, S. 2020. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*.

Gottesman, O.; Johansson, F.; Komorowski, M.; Faisal, A.; Sontag, D.; Doshi-Velez, F.; and Celi, L. A. 2019. Guidelines for reinforcement learning in healthcare. *Nature medicine*, 25(1): 16–18.

He, J.; Zhou, D.; and Gu, Q. 2021. Logarithmic regret for reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, 4171–4180. PMLR.

Hu, J.; Chen, X.; Jin, C.; Li, L.; and Wang, L. 2021. Near-optimal representation learning for linear bandits and linear rl. In *International Conference on Machine Learning*, 4349–4358. PMLR.

Hu, Y.; Kallus, N.; and Uehara, M. 2021. Fast Rates for the Regret of Offline Reinforcement Learning. In Belkin, M.; and Kpotufe, S., eds., *Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado, USA*, volume 134 of *Proceedings of Machine Learning Research*, 2462. PMLR.

Ji, X.; Chen, M.; Wang, M.; and Zhao, T. 2023. Sample Complexity of Nonparametric Off-Policy Evaluation on Low-Dimensional Manifolds using Deep Networks. In *The Eleventh International Conference on Learning Representations*.

Jin, C.; Yang, Z.; Wang, Z.; and Jordan, M. I. 2020. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, 2137–2143. PMLR.

Jin, Y.; Yang, Z.; and Wang, Z. 2021. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, 5084–5096. PMLR.

Kitagawa, T.; and Tetenov, A. 2018. Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice. *Econometrica*, 86(2): 591–616.

Lange, S.; Gabel, T.; and Riedmiller, M. 2012. Batch reinforcement learning. In *Reinforcement learning*, 45–73. Springer.

Lattimore, T.; and Szepesvári, C. 2020. *Bandit Algorithms*. Cambridge University Press.

Le, H. M.; Voloshin, C.; and Yue, Y. 2019. Batch Policy Learning under Constraints. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, 3703–3712. PMLR.

Levine, S.; Kumar, A.; Tucker, G.; and Fu, J. 2020. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*.

Liu, Y.; Swaminathan, A.; Agarwal, A.; and Brunskill, E. 2020. Off-policy policy gradient with stationary distribution correction. In *Uncertainty in Artificial Intelligence*, 1180–1190. PMLR.

Munos, R.; and Szepesvári, C. 2008. Finite-Time Bounds for Fitted Value Iteration. *J. Mach. Learn. Res.*, 9: 815–857.

Nguyen-Tang, T.; and Arora, R. 2023. VIPeR: Provably Efficient Algorithm for Offline RL with Neural Function Approximation. In *The Eleventh International Conference on Learning Representations*.

Nguyen-Tang, T.; Gupta, S.; Nguyen, A. T.; and Venkatesh, S. 2022a. Offline Neural Contextual Bandits: Pessimism, Optimization and Generalization. In *International Conference on Learning Representations*.

Nguyen-Tang, T.; Gupta, S.; Tran-The, H.; and Venkatesh, S. 2022b. On Sample Complexity of Offline Reinforcement Learning with Deep ReLU Networks in Besov Spaces. *Transactions on Machine Learning Research*.

Nie, X.; Brunskill, E.; and Wager, S. 2021. Learning when-to-treat policies. *Journal of the American Statistical Association*, 116(533): 392–409.

Papini, M.; Tirinzoni, A.; Pacchiano, A.; Restelli, M.; Lazaric, A.; and Pirotta, M. 2021. Reinforcement Learning

in Linear MDPs: Constant Regret and Representation Selection. *Advances in Neural Information Processing Systems*, 34.

Simchowitz, M.; and Jamieson, K. G. 2019. Non-asymptotic gap-dependent regret bounds for tabular mdps. *Advances in Neural Information Processing Systems*, 32.

Strehl, A.; Langford, J.; Li, L.; and Kakade, S. M. 2010. Learning from logged implicit exploration data. *Advances in neural information processing systems*, 23.

Szepesvári, C.; and Munos, R. 2005. Finite time bounds for sampling based fitted value iteration. In *Proceedings of the 22nd international conference on Machine learning*, 880–887.

Thomas, P. S.; Theocharous, G.; Ghavamzadeh, M.; Durugkar, I.; and Brunskill, E. 2017. Predictive Off-Policy Policy Evaluation for Nonstationary Decision Problems, with Applications to Digital Marketing. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, 4740–4745. AAAI Press.

Uehara, M.; and Sun, W. 2022. Pessimistic Model-based Offline Reinforcement Learning under Partial Coverage. In *International Conference on Learning Representations*.

Wang, R.; Foster, D.; and Kakade, S. M. 2021. What are the Statistical Limits of Offline RL with Linear Function Approximation? In *International Conference on Learning Representations*.

Wang, X.; Cui, Q.; and Du, S. S. 2022. On Gap-dependent Bounds for Offline Reinforcement Learning. In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho, K., eds., *Advances in Neural Information Processing Systems*.

Xie, T.; Cheng, C.-A.; Jiang, N.; Mineiro, P.; and Agarwal, A. 2021. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34.

Xiong, W.; Zhong, H.; Shi, C.; Shen, C.; Wang, L.; and Zhang, T. 2023. Nearly Minimax Optimal Offline Reinforcement Learning with Linear Function Approximation: Single-Agent MDP and Markov Game. In *The Eleventh International Conference on Learning Representations*.

Xu, H.; Ma, T.; and Du, S. 2021. Fine-grained gap-dependent bounds for tabular mdps via adaptive multi-step bootstrap. In *Conference on Learning Theory*, 4438–4472. PMLR.

Yang, K.; Yang, L.; and Du, S. 2021. Q-learning with logarithmic regret. In *International Conference on Artificial Intelligence and Statistics*, 1576–1584. PMLR.

Yang, L.; and Wang, M. 2019. Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, 6995–7004. PMLR.

Yin, M.; Bai, Y.; and Wang, Y.-X. 2021a. Near-optimal offline reinforcement learning via double variance reduction. *Advances in neural information processing systems*, 34: 7677–7688.

Yin, M.; Bai, Y.; and Wang, Y.-X. 2021b. Near-optimal provable uniform convergence in offline policy evaluation for reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, 1567–1575. PMLR.

Yin, M.; Duan, Y.; Wang, M.; and Wang, Y.-X. 2022. Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism. *International Conference on Learning Representations*.

Yin, M.; and Wang, Y.-X. 2021. Towards instance-optimal offline reinforcement learning with pessimism. *Advances in neural information processing systems*, 34.

Zanette, A.; Wainwright, M. J.; and Brunskill, E. 2021. Provable benefits of actor-critic methods for offline reinforcement learning. *Advances in neural information processing systems*, 34: 13626–13640.

Zhan, R.; Hadad, V.; Hirshberg, D. A.; and Athey, S. 2021a. Off-Policy Evaluation via Adaptive Weighting with Data from Contextual Bandits. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Zhan, R.; Ren, Z.; Athey, S.; and Zhou, Z. 2021b. Policy learning with adaptively collected data. *arXiv preprint arXiv:2105.02344*.

Zhang, K.; Janson, L.; and Murphy, S. 2021. Statistical inference with m-estimators on adaptively collected data. *Advances in Neural Information Processing Systems*, 34: 7460–7471.

Zhang, M.; Nguyen-Tang, T.; Wu, F.; He, Z.; Xie, X.; and Ong, C. S. 2022a. Two-Stage Neural Contextual Bandits for Personalised News Recommendation. *arXiv preprint arXiv:2206.14648*.

Zhang, T.; Ren, T.; Yang, M.; Gonzalez, J.; Schuurmans, D.; and Dai, B. 2022b. Making linear mdps practical via contrastive representation learning. In *International Conference on Machine Learning*, 26447–26466. PMLR.

Zhou, D.; Gu, Q.; and Szepesvari, C. 2021. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, 4532–4576. PMLR.