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

Thanh Nguyen-Tang[1], Ming Yin[2,3], Sunil Gupta[4], Svetha Venkatesh[4], Raman Arora[1]
[1]Department of Computer Science, Johns Hopkins University
[2]Department of Computer Science, UC Santa Barbara
[3]Department of Statistics and Applied Probability, UC Santa Barbara
[4]Applied AI Institute, Deakin University

November 13, 2022

Abstract

Sample-efficient offline reinforcement learning (RL) with linear function approximation has been studied extensively recently. Much of prior work has yielded $\tilde{\mathcal{O}}(\frac{1}{\sqrt{K}})$ sub-optimality bound, which is also minimax optimal, with K being the number of episodes in the offline data. In this work, we seek to understand instance-dependent bounds for offline RL with function approximation. We present an algorithm called Bootstrapped and Constrained Pessimistic Value Iteration (BCP-VI), which leverages data bootstrapping and uses constrained optimization to enforce additional 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 of $\tilde{\mathcal{O}}(\frac{1}{K})$ for offline RL when there is a positive sub-optimality gap in the optimal Q-value function, even when 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 K exceeds a (finite) instance-dependent threshold. To the best of our knowledge, these are the first $\tilde{\mathcal{O}}(\frac{1}{K})$ bound (under a gap assumption) and absolute zero sub-optimality error (additionally under spanning linear features) respectively 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.

1 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 et al., 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 et al., 2021), recommendation systems (Strehl et al., 2010; Thomas et al., 2017; Zhang et al., 2022), econometrics (Kitagawa & Tetenov, 2018; Athey & 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 et al. (2021) present a pessimistic value iteration (PEVI) algorithm and show that it is sample-efficient. In particular, Jin et al. (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. (2022); Yin et al. (2022) leverage variance reduction (to derive a variance-aware bound) and use data-splitting (to circumvent the uniform concentration argument) to further improve the result in Jin et al. (2021) by a factor of $\mathcal{O}(\sqrt{d}H)$. Xiong et al. (2022); Yin et al. (2022) leverage variance reduction (to derive a variance-aware bound) and use data-splitting (to circumvent the uniform concentration argument) to further improve the result in Jin et al. (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 et al. (2021) by a factor of $\sqrt{d}$ when the action space is finite, and the function approximation is linear. Uehara & Sun (2021) also obtain a convergence rate of $\frac{1}{\sqrt{K}}$ 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 convergence rate of $\frac{1}{\sqrt{K}}$ is minimax-optimal for offline RL, 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). 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, existing works establishing the standard convergence rate of $\frac{1}{\sqrt{K}}$ still require a strong assumption of uniform feature coverage and trajectory independence. [1] This motivates us to study tighter instance-dependent bounds for offline RL with the mildest data coverage condition possible.

Instance- or gap-dependent bounds have been extensively studied in *online* bandit and reinforcement learning literature (Simchowitz & Jamieson, 2019; Yang et al., 2021; Xu et al., 2021; He et al., 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 et al., 2021; Wang et al., 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, these works (Hu et al., 2021; Wang et al., 2022) 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-dependent/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 was discussed in the concurrent work of Wang et al. (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 & Jamieson, 2019; Yang et al., 2021; He et al., 2021). The larger the $\Delta_{\min}$, the faster we can learn in an online setting since the actions with larger rewards are

[1]The only exception is Jin et al. (2021), but their bound is generic, and they do not show if they can achieve a rate of $\frac{1}{\sqrt{K}}$ under a partial data coverage assumption.

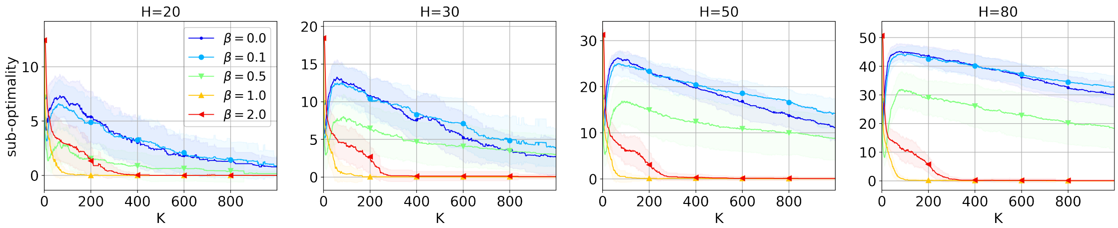

Figure 1: A comparison of BCP-VI with its non-pessimistic variant (i.e., $\beta = 0$). The plots show the sub-optimality of $\hat{\pi}^K$ returned by Algorithm 1 for $K \in [1, \ldots, 1000]$ and various values of episode lengths $H \in \{20, 30, 50, 80\}$. The hyperparameter β is the pessimism parameter given on line 7 of Algorithm 1.

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 et al., 2021). However, it remains unclear *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.*

1.1 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 et al. (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 the uniform mixture policy from the ensemble learned by the BCP-VI adapts to the instance-dependent quantity, Δ_{min}, to achieve a fast rate of $\mathcal{O}(\frac{\log K}{K})$ where K is the number of episodes ihttps://www.overleaf.com/projectn the offline data. Our result holds under the single-policy concentration coverage even when the offline data are collected adaptively. This fills the gap in the literature about instance/gap-dependent bounds for offline RL with linear function approximation and adaptive data.

2. As a byproduct, we also derive strong 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 (and without the gap information). The previous works in offline linear MDPs (Jin et al., 2021; Xie et al., 2021) only show how to turn their learning bounds into a $\frac{1}{\sqrt{K}}$ bound with the uniform data feature coverage assumption but not with the optimal-policy concentrability.

3. Under an additional condition (besides the gap assumption) 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 zero sub-optimality and thus are provably optimal.

4. We accompany our main result with information-theoretic lower bounds, which show

Algorithm	Condition	Upper Bound	Lower Bound	Data
PEVI	Uniform	$\tilde{\mathcal{O}}\left(\frac{H^2 d^{3/2}}{\sqrt{K}}\right)$	$\Omega\left(\frac{H}{\sqrt{K}}\right)$	Independent
BCP-VI	OPC	$\tilde{\mathcal{O}}\left(\frac{H d^{3/2}\kappa}{\sqrt{K}}\right)$	$\Omega\left(\frac{H\sqrt{\kappa_{\min}}}{\sqrt{K}}\right)$	Adaptive
	OPC, Δ_{min}	$\tilde{\mathcal{O}}\left(\frac{d^3 H^2 \kappa^3}{\Delta_{\min}\cdot K}\right)$	$\Omega\left(\frac{H^2\kappa_{\min}}{\Delta_{\min}\cdot K}\right)$	Adaptive
	OPC, Δ_{min}, UO-SF	0 if $K \geq k^*$ defined in Eq. (2)	NA	Adaptive
BCP-VTR	OPC	$\tilde{\mathcal{O}}\left(\frac{H d\kappa}{\sqrt{K}}\right)$	$\Omega\left(\frac{H\sqrt{\kappa_{\min}}}{\sqrt{K}}\right)$	Adaptive
	OPC, Δ_{min}	$\tilde{\mathcal{O}}\left(\frac{d^2 H^2 \kappa^3}{\Delta_{\min}\cdot K}\right)$	$\Omega\left(\frac{H^2\kappa_{\min}}{\Delta_{\min}\cdot K}\right)$	Adaptive

Table 1: Bounds on the sub-optimality of offline RL with linear function approximation under different conditions and data coverage assumptions. Cells in gray are our contributions. The results in the first line were obtained in Jin et al. (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 4.1), $\kappa = \sum_{h=1}^{H} \kappa_h$ where κ_h is the OPC coefficient defined in Assumption 4.2, $\kappa_{\min} = \min_{h\in[H]} \kappa_h$, "Uniform" means uniform data coverage, "Independent" and "Adaptive" mean the episodes of the offline data are collected independently and adaptively, respectively, and UO-SF stands for unique optimality and spanning features in Assumption 4.4. BCP-VTR is a model-based offline RL method for linear mixture MDPs which is presented in Section F.

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.

2 Related Work

Offline RL with linear function approximation. While there has been much focus on provably efficient RL under linear function approximation, Jin et al. (2021) were the first to show that pessimistic value iteration is provably efficient for offline linear MDPs. Xiong et al. (2022) and Yin et al. (2022) improve upon Jin et al. (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 et al. (2021) by a factor of $\sqrt{d}$ when realized to finite action spaces and linear MDPs. However, all of the results above yield a worst-case bound of $\frac{1}{\sqrt{K}}$ without taking into account the structure of a particular problem instance. On the other hand, Wang et al. (2020) 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.

Instance-dependent bounds for offline RL. The gap assumption (Assumption 4.3) has been studied extensively in online RL (Bubeck & Cesa-Bianchi, 2012; Lattimore & Szepesvári, 2020), yielding gap-dependent logarithmic regret bounds for bandits, tabular MDPs (Yang et al., 2021) and MDPs with linear representation (He et al., 2021). In online RL, when learning MDPs with linear rewards, under an additional assumption that the linear features of optimal actions span the space of the linear features of all actions (Papini et al., 2021a), 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 et al. (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. *[**Raman**: Confused by the discussion of the rates in the sentence that follows. Can you please clarify?]* Hu et al. (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 et al. (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., 2022; Zhang et al., 2021). Pessimistic value iteration (PEVI) (Jin et al., 2021) works in linear MDP for the general data compliance assumption (see (Jin et al., 2021, Definition 2.1)), which is essentially equivalent to assuming that the data is collected adaptively. *[**Raman**: I am not sure what you mean by the sentence that follows:]* However, when deriving the explicit $\frac{1}{\sqrt{K}}$ bound of their algorithm, they resort to the assumption that the trajectories are independent (see their Corollary 4.6); we argue that it is not necessary. The recent work of Wang et al. (2022) derives a gap-dependent bound for offline tabular MDP but still requires that trajectories are collected independently.

3 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. A time-inhomogeneous transition kernel $\mathbb{P} = \{\mathbb{P}_h\}_{h=1}^H$, where $\mathbb{P}_h : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$ (where $\mathcal{P}(\mathcal{S})$ denotes the set of probability measures over $\mathcal{S}$) maps each state-action pair (s_h, a_h) to a probability distribution $\mathbb{P}_h(\cdot|s_h, a_h)$ (with the corresponding density function $p_h(\cdot|s_h, a_h)$ with respect to the Lebesgue measure ρ on $\mathcal{S}$), and $r = \{r_h\}_{h=1}^H$ 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 $(P_h V)(s, a) := \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s,a)}[V(s')]$, $(T_h V)(s, a) := r_h(s, a) + (P_h V)(s, a)$, 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 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 $\mathcal{SA}^{\mathcal{M},\pi}$ to become 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 $d_h^\pi(s_h, a_h) \leq 1, \forall (h, s_h, a_h, \pi)$.[2]

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 & 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 $\|\theta_h\|_2 \leq \sqrt{d}$, $\|\int \mu_h(s)v(s)ds\|_2 \leq \sqrt{d}\|v\|_\infty$ for any $v : \mathcal{S} \to \mathbb{R}$ and $\|\phi_h(s,a)\|_2 \leq 1$.

Remark 1. *We only consider linear MDP in the main paper but also consider a linear mixture model (Cai et al., 2020; Zhou et al., 2021) in Section F.*

Offline Regime. In the offline learning setting, the goal is to learn the policy π that maximizes v^π, given the 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:

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

Performance metric. We measure the performance of policy $\hat{\pi}$ via the sub-optimality metric: $\text{SubOpt}(\hat{\pi}) := \mathbb{E}_{s_1 \sim d_1}[\text{SubOpt}(\hat{\pi}; s_1)]$, where $\text{SubOpt}(\hat{\pi}; s) := V_1^{\pi^*}(s_1) - V_1^{\hat{\pi}}(s_1)$. As $\hat{\pi}$ is learnt from the offline data $\mathcal{D}$, $\text{SubOpt}(\pi)$ is random (with respect to the randomness of $\mathcal{D}$ and possibly 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.

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

4 Bootstrapped and Constrained Pessimistic Value Iteration

In this section, we describe our main algorithm and establish both instance-agnostic and instance-dependent bounds for offline RL from adaptive data with linear function approximation. Through this algorithm, we show that offline RL achieves a generic data-dependent bound under the optimal-policy concentrability and adapts to the gap information to accelerate to $\frac{\log K}{K}$-type bound and even obtain zero sub-optimality when the optimal linear features under the behavior policy spans those under an optimal policy.

4.1 Algorithm

We build upon the Pessimistic Value Iteration (PEVI) algorithm (Jin et al., 2021) with two additional modifications: bootstrapping and constrained optimization, thus the name Bootstrapped and Constrained Pessimistic Value Iteration (BCP-VI) in Algorithm 1. The constrained optimization in 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).[3] The additional modifications are highlighted in blue in Algorithm 1.

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 $\mathcal{D}^{k-1}$:

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

In 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 pessimitic estimate of the value functions. In Line 10, we extract policy $\hat{\pi}_h$ that is most greedy with respect to $\hat{Q}_h$ 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 policies from the ensemble as the execution policy: the *mixture* policy $\hat{\pi}^{mix}$ and the *last-iteration* policy $\hat{\pi}^{last}$, defined as: $\hat{\pi}^{mix} := \frac{1}{K} \sum_{k=1}^{K} \hat{\pi}^k$, and $\hat{\pi}^{last} := \hat{\pi}^{K+1}$. Note that $\hat{\pi}^{last}$ is similar to the PEVI policy in Jin et al. (2021).

4.2 Data-Dependent Bound

Sample-efficient offline reinforcement learning is not possible without certain data-coverage assumptions (Wang et al., 2020). In this work, we rely on the optimal-policy concentrability (Assumption 4.1) which ensures that d^μ covers the trajectory of some optimal policy π^* and can be agnostic to other locations.

Assumption 4.1 (Optimal-Policy Concentrability (OPC) (Liu et al., 2019))**.** *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 4.1 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 $\pi_h^* \in \Pi_h(\mu), \forall h \in [H]$.*

[3]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)

```
Input: Dataset D = {(s_h^t, a_h^t, r_h^t)}_{h∈[H]}^{t∈[K]}, uncertainty parameters {β_k}_{k∈[K]}, regularization hyperparameter λ, μ-supported policy class {Π_h(μ)}_{h∈[H]}.
for k = 1, ..., K + 1 do
    V̂_{H+1}^k(·) ← 0.
    for step h = H, H − 1, ..., 1 do
        Σ_h^k ← Σ_{t=1}^{k−1} φ_h(s_h^t, a_h^t) · φ_h(s_h^t, a_h^t)^T + λ · I.
        ŵ_h^k ← (Σ_h^k)^{−1} Σ_{t=1}^{k−1} φ_h(s_h^t, a_h^t) · (r_h^t + V̂_{h+1}^k(s_{h+1}^t)).
        b_h^k(·,·) ← β_k · ‖φ_h(·,·)‖_{(Σ_h^k)^{−1}}.
        Q̄_h^k(·,·) ← ⟨φ_h(·,·), ŵ_h^k⟩ − b_h^k(·,·).
        Q̂_h^k(·,·) ← min{Q̄_h^k(·,·), H − h + 1}^+.
        π̂_h^k ← arg max_{π_h∈Π_h(μ)} ⟨Q̂_h^k, π_h⟩
        V̂_h^k(·) ← ⟨Q̂_h^k(·,·), π_h^k(·|·)⟩.
    end for
end for
Output: Ensemble {π̂^k : k ∈ [K + 1]}.
```

Assumption 4.1 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 Appendix A.5 for a proof that the OPC condition is necessary). As such, Assumption 4.1 is significantly weaker than *uniform* data coverage assumption which features in most existing works in offline RL e.g., the uniform feature coverage (Duan et al., 2020; Yin et al., 2022), i.e., 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 $\min_{h,s_h,a_h} d_h^\mu(s_h,a_h) > 0$, and the classical uniform concentrability (Szepesvári & Munos, 2005; Chen & Jiang, 2019; Nguyen-Tang et al., 2021), i.e., $\sup_{\pi,h,s_h,a_h} \frac{d_h^\pi(s_h,a_h)}{d_h^\mu(s_h,a_h)} < \infty$.

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

Assumption 4.2. $\kappa_h^{-1} := \inf_{(s_h,a_h):d_h^\mu(s_h,a_h)>0} d_h^\mu(s_h,a_h) > 0, \forall h \in [H]$.

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 (but can be exponentially large). We remark that Assumption 4.2 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 et al., 2021) as the infimum in the latter is uniformly over all states and actions. Note that Assumption 4.2 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 4.1, κ_h can be seen as (an upper bound on) the *OPC coefficient* at stage h as we have $\frac{d_h^{\pi^*}(s_h,a_h)}{d_h^\mu(s_h,a_h)} \leq \kappa_h, \forall (h, s_h, a_h) \in [H] \times \mathcal{S} \times \mathcal{A}$.

We fix any $\delta \in (0,1]$ and 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. We now present our data-dependent bound.

Theorem 1 (Data-dependent bound). *Under Assumption 4.1, with probability at least $1 - 4\delta$ over the randomness of $\mathcal{D}$, we have:*

$$\mathrm{SubOpt}(\hat{\pi}^{mix}) \vee \mathrm{SubOpt}(\hat{\pi}^{last}) \leq \frac{4\beta(\delta)}{K} \sum_{h=1}^{H} \sum_{k=1}^{K} \frac{d_h^*(s_h^k, a_h^k)}{d_h^\mu(s_h^k, a_h^k)} \|\phi_h(s_h^k, a_h^k)\|_{(\Sigma_h^k)^{-1}}$$

$$+ \frac{4\beta(\delta)}{K} \sum_{h=1}^{H} \sqrt{\log\left(\frac{H}{\delta}\right) \sum_{k=1}^{K} \left(\frac{d_h^*(s_h^k, a_h^k)}{d_h^\mu(s_h^k, a_h^k)}\right)^2} + \frac{2}{K} + \frac{16H}{3K} \log\left(\frac{\log_2(KH)}{\delta}\right).$$

Remark 3. *The first term in the bound in Theorem 1 is the elliptical potential that emerges from pessimism and is dominant while the other terms are generalization errors by concentration phenomenon and 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 4.1). One immediate consequence of the data-dependent bound in Theorem 1 is that the bound can turn into a weaker yet more explicit $\frac{1}{\sqrt{K}}$ bound in Corollary 1.

Corollary 1. *Under Assumptions 4.1-4.2, with probability at least $1 - \Omega(1/K)$ over the randomness of $\mathcal{D}$, we have:*

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

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

As the $\frac{1}{\sqrt{K}}$ bound is minimax, we compare our result to other existing works.

Comparing with (Yin & Wang, 2021). Yin & 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 et al., 2021). Similarly, Jin et al. (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 et al., 2020; Yin & 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 does not need pessimism to obtain such a bound and Duan et al. (2020); Yin & Wang (2021) do not answer the question of how to obtain the $\frac{1}{\sqrt{K}}$ bound with only OPC.

Comparing with (Xie et al., 2021; Zhan et al., 2022; Chen & Jiang, 2022). Xie et al. (2022); Zhan et al. (2022); Chen & Jiang (2022) consider offline RL with general function approximation under OPC. While their results are valuable for the offline RL literature, their algorithms are more complex and, in general, intractable, in practice. When realized to linear MDPs, the algorithm in Xie et al. (2021) is computationally tractable but their guarantee requires the behaviour policy to be explorative all dimensions of the feature mapping i.e. $\mathbb{E}_\mu[\phi(s,a)\phi(s,a)^T]$ is strictly positive definite.

Theorem 1 is a byproduct that sets the stage for our instance-dependent bounds in the following section; nonetheless, it is the first, to the best of our knowledge, that provides an explicit $\frac{1}{\sqrt{K}}$ bound for linear MDPs with OPC.

Remark 4. *We show in Appendix A.5 that OPC is necessary to guarantee sublinear sub-optimality bound for offline RL. When OPC fails to hold, we show in Appendix A.4 that BCPVI suffers a constant sub-optimality incurred at optimal locations that are not supported by the behavior policy. By intuition, such a constant sub-optimality occurs due to the off-support actions and vanishes when the behaviour policy covers the trajectories of at least one optimal policy (Assumption 4.1).*

4.3 Instance-Dependent Bounds

We now show that offline RL can exploit various types of instance-dependent structures of the underlying MDP to speed up the sub-optimality rate.

4.3.1 Gap-Dependent Bound

The first measure of the hardness of an MDP instance is the minimum positive action gap (Assumption 4.3) 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:*

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

In this paper, we assume that the minimal sub-optimality gap is strictly positive, which is a common assumption for gap-dependent analysis (Simchowitz & Jamieson, 2019; Yang et al., 2021; He et al., 2021).

Assumption 4.3 (Minimum positive sub-optimality gap)**.** $\Delta_{\min} > 0$.

We now present the sub-optimality bound under the gap information.

Theorem 2 ($\frac{\log K}{K}$ sub-optimality bound)**.** *Under Assumptions 4.1-4.2-4.3, with probability at least $1 - (1 + 3 \log_2(H/\Delta_{\min}))\delta$,*

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

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

$$\mathbb{E}\left[\mathrm{SubOpt}(\hat{\pi}^{mix})\right] = \tilde{\mathcal{O}}\left(\frac{d^3 H^2 \kappa^3}{\Delta_{\min} \cdot K}\right).$$

The sub-optimality 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. To our knowledge, this is the first theoretical result that can leverage the gap information $\Delta_{\min}$ to accelerate to a $\mathcal{O}\left(\frac{\log K}{K}\right)$ 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 4.1-4.2-4.3.

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:*

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

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.

4.3.2 Absolute Zero Sub-optimality

We introduce extra assumptions on the linear mapping which our algorithm can exploit further to accelerate the rate. For that, we assume the unique optimality and the spanning feature property.

Assumption 4.4 (Unique Optimality and Spanning features)**.** *We assume that*

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

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

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

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

where $span(\mathcal{X})$ denotes the vector space spanned by all linear combination of elements in $\mathcal{X}$.

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 4.4.2 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., 2021a). 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 4.4.2 instead requires such condition only over optimal actions and states reachable by the behavior policy. Note that Assumption 4.4.2 also does not require the linear feature under d_h^* to span the full space $\mathbb{R}^d$.

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. *Under Assumption 4.1-4.2-4.3-4.4, then with probability at least $1 - 4\delta$, we have:* $\text{SubOpt}(\hat{\pi}^k) = 0, \forall k \geq k^*$, *where k^* is defined in Eq. (2).*

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

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 a instance-dependent threshold specified by k^*, the policies $\hat{\pi}^k$ in the ensemble precisely match the (unique) optimal policy with high probability. This is a first theoretical result for constant bound of offline RL with linear function approximation under partial data coverage and adaptive data.

5 Proof overview

In the following, we provide a brief overview of the key proof ideas. The details proofs are deferred to the appendix.

For Theorem 1. With the extended value difference and the constrained optimization in Line 10 of Algorithm 1, we can convert bounding $\mathrm{SubOpt}(\hat{\pi}^k)$ 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

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

since $\Sigma_h^k \preceq \Sigma_h^{K+1}$.

For Theorem 2. We convert bounding $\mathrm{SubOpt}(\hat{\pi}^{mix})$ to bounding its empirical quantity $\frac{1}{K}\sum_{k=1}^K \mathrm{SubOpt}(\hat{\pi}^k; s_1^k)$ plus a generalization term. Using the original online-to-batch argument (Cesa-Bianchi et al., 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 A.5) with $\tilde{\mathcal{O}}(\frac{1}{K})$ generalization error which is of independent interest. Then, $\mathrm{SubOpt}(\hat{\pi}^k; s_1)$ is expressed through decomposition $\mathrm{SubOpt}(\hat{\pi}^k; s_1) = \mathbb{E}_{\hat{\pi}^k}[\sum_{h=1}^H \Delta_h(s_h, a_h)|\mathcal{F}_{k-1}, s_1]$ (Lemma B.1). To handle the gap terms, the key observation is that $\hat{\pi}^k$ belongs to the μ-supported policy class (Lemma B.2), thus the concentrability coefficients (Assumption 4.2) 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, $\sum_{k=1}^K \mathbb{1}\{\Delta_h(s_h^k, \hat{\pi}_h^k(s_h^k)) \ge \Delta\} \lesssim \frac{d^3H^2\iota^{-2}}{\Delta^2}\log^3(dKH/\delta)$ (Lemma B.4).

For Theorem 4. A key observation is that $\lambda_{\min}(\Sigma_h^k) \gtrsim k\lambda_h^+$ (Lemma C.2) where λ_h^+ is the minimum positive eigenvalue of Σ_h^*. Thus, for any $v \in \mathrm{span}(\{\phi_h^*(s)|s \in \mathcal{S}_h^*\})$, we have $\|v\|_{(\Sigma_h^k)^{-1}} \le \mathcal{O}(1/\sqrt{k})$ (Lemma C.3). With Assumption 4.4, $\forall s_h \in \mathcal{S}_h^\mu$, $\Delta_h(s_h, \hat{\pi}_h^k(s_h)) \le 2\beta_k\mathbb{E}_{\pi^*}[\sum_{h'=h}^H \|\phi_h(s_h, a_h)\|_{(\Sigma_{h'}^k)^{-1}}|\mathcal{F}_{k-1}, s_h] = \mathcal{O}(\frac{1}{\sqrt{k}})$ and $\mathcal{O}(\frac{1}{\sqrt{k}}) < \Delta_{\min}$, for sufficiently large k.

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

6 Discussion

This work studies offline RL with linear function approximation and contributes the first $\tilde{\mathcal{O}}(\frac{1}{K\Delta_{\min}})$ bound and constant bound in this setting, using bootstrapping and constrained optimization on top of pessimism.

One open question is that there is still a gap between upper bounds and lower bounds in terms of d and κ. While the gap for d can be potentially tightened if we leverage the variance information (Yin et al., 2022) (at the potential expense of a stronger assumption in data coverage), it is open how to close the gap in κ in linear MDPs.

References

Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper/2011/file/e1d5be1c7f2f456670de3d53c7b54f4a-Paper.pdf.

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

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pp. 463–474. PMLR, 2020.

Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

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

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pp. 1283–1294. PMLR, 2020.

Nicolò Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the generalization ability of on-line learning algorithms. *IEEE Trans. Inf. Theory*, 50(9):2050–2057, 2004. doi: 10.1109/TIT.2004.833339. URL https://doi.org/10.1109/TIT.2004.833339.

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

Jinglin Chen and Nan Jiang. Offline reinforcement learning under value and density-ratio realizability: the power of gaps. 2022.

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

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. Guidelines for reinforcement learning in healthcare. *Nature medicine*, 25(1):16–18, 2019.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Logarithmic regret for reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pp. 4171–4180. PMLR, 2021.

Yichun Hu, Nathan Kallus, and Masatoshi Uehara. Fast rates for the regret of offline reinforcement learning. *arXiv preprint arXiv:2102.00479*, 2021.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143. PMLR, 2020.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pp. 5084–5096. PMLR, 2021.

Toru Kitagawa and Aleksey Tetenov. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018. doi: https://doi.org/10.3982/ECTA13288. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA13288.

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

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. doi: 10.1017/9781108571401.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Off-policy policy gradient with stationary distribution correction. In Amir Globerson and Ricardo Silva (eds.), *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019, Tel Aviv, Israel, July 22-25, 2019*, volume 115 of *Proceedings of Machine Learning Research*, pp. 1180–1190. AUAI Press, 2019. URL http://proceedings.mlr.press/v115/liu20a.html.

Yifei Min, Tianhao Wang, Dongruo Zhou, and Quanquan Gu. Variance-aware off-policy evaluation with linear function approximation. *Advances in neural information processing systems*, 34, 2021.

Thanh Nguyen-Tang, Sunil Gupta, Hung Tran-The, and Svetha Venkatesh. Sample complexity of offline reinforcement learning with deep relu networks, 2021.

Thanh Nguyen-Tang, Sunil Gupta, A. Tuan Nguyen, and Svetha Venkatesh. Offline neural contextual bandits: Pessimism, optimization and generalization. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=sPIFuucA3F.

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

Matteo Papini, Andrea Tirinzoni, Aldo Pacchiano, Marcello Restelli, Alessandro Lazaric, and Matteo Pirotta. Reinforcement learning in linear mdps: Constant regret and representation selection. *Advances in Neural Information Processing Systems*, 34, 2021a.

Matteo Papini, Andrea Tirinzoni, Marcello Restelli, Alessandro Lazaric, and Matteo Pirotta. Leveraging good representations in linear contextual bandits. *CoRR*, abs/2104.03781, 2021b. URL https://arxiv.org/abs/2104.03781.

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

Alex Strehl, John Langford, Sham Kakade, and Lihong Li. Learning from logged implicit exploration data. *arXiv preprint arXiv:1003.0120*, 2010.

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

Philip S. Thomas, Georgios Theocharous, Mohammad Ghavamzadeh, Ishan Durugkar, and Emma Brunskill. 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, pp. 4740–4745. AAAI Press, 2017.

Joel Tropp. Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270, 2011.

Masatoshi Uehara and Wen Sun. Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*, 2021.

Ruosong Wang, Dean P Foster, and Sham M Kakade. What are the statistical limits of offline rl with linear function approximation? *arXiv preprint arXiv:2010.11895*, 2020.

Xinqi Wang, Qiwen Cui, and Simon S Du. On gap-dependent bounds for offline reinforcement learning. *arXiv preprint arXiv:2206.00177*, 2022.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34, 2021.

Wei Xiong, Han Zhong, Chengshuai Shi, Cong Shen, Liwei Wang, and T. Zhang. Nearly minimax optimal offline reinforcement learning with linear function approximation: Single-agent mdp and markov game. *ArXiv*, abs/2205.15512, 2022.

Haike Xu, Tengyu Ma, and Simon Du. Fine-grained gap-dependent bounds for tabular mdps via adaptive multi-step bootstrap. In *Conference on Learning Theory*, pp. 4438–4472. PMLR, 2021.

Kunhe Yang, Lin Yang, and Simon Du. Q-learning with logarithmic regret. In *International Conference on Artificial Intelligence and Statistics*, pp. 1576–1584. PMLR, 2021.

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

Ming Yin and Yu-Xiang Wang. Towards instance-optimal offline reinforcement learning with pessimism. *Advances in neural information processing systems*, 34, 2021.

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*, pp. 1567–1575. PMLR, 2021.

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

Ruohan Zhan, Vitor Hadad, David A. Hirshberg, and Susan Athey. Off-policy evaluation via adaptive weighting with data from contextual bandits. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Aug 2021a. doi: 10.1145/3447548.3467456. URL http://dx.doi.org/10.1145/3447548.3467456.

Ruohan Zhan, Zhimei Ren, Susan Athey, and Zhengyuan Zhou. Policy learning with adaptively collected data. *arXiv preprint arXiv:2105.02344*, 2021b.

Wenhao Zhan, Baihe Huang, Audrey Huang, Nan Jiang, and Jason D Lee. Offline reinforcement learning with realizability and single-policy concentrability. *arXiv preprint arXiv:2202.04634*, 2022.

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

Mengyan Zhang, Thanh Nguyen-Tang, Fangzhao Wu, Zhenyu He, Xing Xie, and Cheng Soon Ong. Two-stage neural contextual bandits for personalised news recommendation. *arXiv preprint arXiv:2206.14648*, 2022.

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pp. 4532–4576. PMLR, 2021.

Notations	Meaning
$\mathcal{M}$	MDP instance, $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbb{P}, r, H, d_1)$
$\mathcal{S}$	The arbitrary state space in $\mathbb{R}^d$
$\mathcal{A}$	The arbitrary action space
H	Episode length
K	Number of episodes
$\mathbb{P}_h(\cdot\|s_h, a_h)$	Next state probability distribution
$p_h(\cdot\|s_h, a_h)$	Next state density function (with respect to the Lebesgue measure)
$r_h(s_h, a_h)$	Reward function in $[0, 1]$
$P_h V$	$(P_h V)(s, a) = \mathbb{E}_{s' \sim \mathbb{P}_h(\cdot\|s,a)}\left[V(s')\right]$
$T_h V$	$r_h + P_h V$
$\pi^* = \{\pi_h^*\}_{h\in[H]}$	Optimal policy
$Q^\pi = \{Q_h^\pi\}_{h\in[H]}$	Action-value functions under policy π
$V^\pi = \{V_h^\pi\}_{h\in[H]}$	Value functions under policy π
d_h^π	Marginal state-visitation density (with respect to the Lebesgue measure)
d_1	Initial state density
d_h^*	$d_h^{\pi^*}$
$\phi_h(s_h, a_h)$	Feature at step h
$\phi_h^*(s_h)$	$\phi_h(s_h, a_h^*)$ where $a_h^* \sim \pi_h^*(\cdot\|s_h)$ (thus $\phi_h^*(s_h)$ is random w.r.t. π_h^*)
Σ_h^*	$\mathbb{E}_{s_h \sim d_h^*}\left[\phi_h^*(s_h)\phi_h^*(s_h)^T\right]$
$\Delta_h(s_h, a_h)$	$V_h^*(s_h) - Q^*(s_h, a_h)$
$\Delta_{\min}$	$\min_{h,s_h,a_h} \Delta_h(s_h, a_h)$
μ	Behavior policy
κ_h	$\max_{s,a} \frac{d_h^*(s,a)}{d_h^\mu(s,a)}$
κ	$\sum_{h=1}^H \kappa_h$
$\kappa_{1:h}$	$\prod_{i=1}^h \kappa_h$
τ_h	$\{(s_i, a_i, r_i)\}_{i\in[h]}$
λ	Regularization parameter
$\mathcal{D}$	The offline data, $\{(s_h^k, a_h^k, r_h^k)\}_{h\in[H]}^{k\in[K]}$
$\mathcal{F}_k$	$\sigma(\{(s_h^t, a_h^t, r_h^t)\}_{h\in[H]}^{t\in[k]})$ where $\sigma(\cdot)$ is the σ-algebra generated by $(\cdot)$
$\mathrm{SubOpt}(\pi; s)$	$V_1^*(s) - V^\pi(s)$
$\mathrm{SubOpt}(\pi)$	$\mathbb{E}_{s\sim d_1}\left[\mathrm{SubOpt}(\pi; s)\right]$
$\hat{\pi}_h^{mix}(a_h\|s_h)$	$\frac{1}{K}\sum_{k=1}^K \hat{\pi}_h^k(a_h\|s_h)$
$\hat{\pi}_h^{last}(a_h\|s_h)$	$\hat{\pi}^{K+1}$
δ	Failure probability
$\beta_k(\delta)$	$C \cdot dH \log(dHk/\delta)$
$\\|A\\|_2$	The spectral norm of matrix A, i.e. $\lambda_{\max}(A)$
$A \succeq B$	$A - B$ is positive semi-definite (p.s.d.)
$\mathrm{supp}(p)$	The support set of density p, i.e. $\{s : p(s) > 0\}$
$\\|v\\|_2$	$\sqrt{\sum_{i=1}^d v_i^2}$
$\\|v\\|_\infty$	$\max_{i\in[d]} v_i$
WLOG	Without loss of generality
$\mathrm{poly}\log K$	A polynomial of $\log K$
$\mathcal{O}(\cdot)$	Big-Oh notation
$\tilde{\mathcal{O}}(\cdot)$	Big-Oh notation with hidden log factors
$\Omega(\cdot)$	Big-Omega notation
$\tilde{\Omega}(\cdot)$	Big-Omega notation with hidden log factors

Table 2: Notations

A Proof of The Results In Section 4.2

In this section, we provide the detailed proofs for all the results stating in Section 4.2, including Theorem 1, Corollary 1, and Remark 4. For convenience, we present all notations in Table 2.

A.1 Proof for the mixture policy in Theorem 1

We first provide the proof for the bound of $\hat{\pi}^{mix}$ in Theorem 1. We decompose the proof into three main steps.

Step 1: Bounding sub-optimality of each $\hat{\pi}^k$. We first construct pointwise bounds for the sub-optimality of each bootstrapped policies $\hat{\pi}^k$ to bound $\text{SubOpt}(\hat{\pi}^k; s), \forall (k, s) \in [K] \times \mathcal{S}$. The techniques for this step are quite standard that use pessimism, linear representation and self-normalized martingale concentrations. Define the Bellman error

$$\zeta_i^k(s,a) := (T_i \hat{V}_{i+1}^k)(s,a) - \hat{Q}_i^k(s,a), \forall (i,k,s,a) \in [H] \times [K] \times \mathcal{S} \times \mathcal{A}.$$

We show that if the uncertainty quantifier function b_h^k bounds the error of the empirical Bellman operator, then the Bellman error is non-negative and is bounded above by a constant factor of the uncertainty quantifier function.

Lemma A.1. *$\forall (h,k,s,a) \in [H] \times [K] \times \mathcal{S} \times \mathcal{A}$, if $|(T_h \hat{V}_{h+1}^k)(s,a) - (\hat{T}_h^k \hat{V}_{h+1}^k)(s,a)| \leq b_h^k(s,a)$, then $\zeta_h^k(s,a) \leq 2 b_h^k(s,a)$.*

Proof of Lemma A.1. We fix any $(h,k,s,a) \in [H] \times [K] \times \mathcal{S} \times \mathcal{A}$. First, we show that: $\zeta_h^k(s,a) \geq 0$. Indeed, if $\bar{Q}_i^k(s,a) < 0$, $\hat{Q}_i^k(s,a) = 0$, thus $\zeta_i^k(s,a) = (T_i \hat{V}_{i+1}^k)(s,a) - \hat{Q}_i^k(s,a) = (T_i \hat{V}_{i+1}^k)(s,a) \geq 0$. If $\bar{Q}_i^k(s,a) \geq 0$, we have:

$$\begin{aligned}
\zeta_i^k(s,a) &= (T_i \hat{V}_{i+1}^k)(s,a) - \hat{Q}_i^k(s,a) \geq (T_i \hat{V}_{i+1}^k)(s,a) - \bar{Q}_i^k(s,a) \\
&= (T_i \hat{V}_{i+1}^k)(s,a) - (\hat{T}_h^k \hat{V}_{h+1}^k)(s,a) + b_t^k(s,a) \geq 0.
\end{aligned}$$

Next we show that: $\zeta_h^k(s,a) \leq 2 b_h^k(s,a)$. We have $\bar{Q}_h^k(s,a) = (\hat{T}_h^k \hat{V}_{h+1}^k)(s,a) - b_h^k(s,a) \leq (\hat{T}_h^k \hat{V}_{h+1}^k)(s,a) \leq H - h + 1$. Thus, $\hat{Q}_i^k(s,a) = \max\{\bar{Q}_i^k(s,a), 0\}$. Thus, we have:

$$\begin{aligned}
\zeta_i^k(s,a) &= (T_i \hat{V}_{i+1}^k)(s,a) - \hat{Q}_i^k(s,a) \leq (T_i \hat{V}_{i+1}^k)(s,a) - \bar{Q}_i^k(s,a) \\
&= (T_i \hat{V}_{i+1}^k)(s,a) - (\hat{T}_h^k \hat{V}_{h+1}^k)(s,a) + (\hat{T}_h^k \hat{V}_{h+1}^k)(s,a) - \bar{Q}_i^k(s,a) \leq 2 b_h^k(s,a).
\end{aligned}$$

□

We show that the stage-wise sub-optimality is bounded by the sum of the uncertainty quantifier functions along the trajectories of the optimal policy.

Lemma A.2. *Suppose that w.p.a.l. $1-\delta$: $|(T_h \hat{V}_{h+1}^k)(s,a) - (\hat{T}_h^k \hat{V}_{h+1}^k)(s,a)| \leq b_h^k(s,a), \forall (h,k,s,a) \in [H] \times [K] \times \mathcal{S} \times \mathcal{A}$. Then, under Assumption 4.1, w.p.a.l. $1-\delta$, we have:*

$$\forall (h, s_h, k) \in \mathcal{S} \times [H] \times [K], V_h^{\pi^*}(s_h) - V_h^{\hat{\pi}^k}(s_h) \leq 2 \mathbb{E}_{\pi^*} \left[\sum_{i=h}^{H} b_i^k(s_i, a_i) | s_h \right].$$

Proof of Lemma A.2. Consider any $(s_h, a_h) \in \mathcal{SA}_h^{\pi^*}$ (i.e. $d_h^*(s_h, a_h) > 0$). Recall from Line 10 of Algorithm 1 that $\hat{\pi}_i^k = \arg\max_{\pi_i \in \Pi_i(\mu)} \langle \hat{Q}_i^k, \pi_i \rangle, \forall i$, and from Remark 2 (under Assumption 4.1), that $\pi_i^* \in \Pi_i(\mu), \forall i \in [H]$. Thus, we have:

$$\langle \hat{Q}_i^k, \hat{\pi}_i^k \rangle \geq \langle \hat{Q}_i^k, \pi_i^* \rangle, \forall i \in [H].$$

Then, by the value decomposition lemma (Lemma E.2), $\forall(h, s_h)$, we have

$$
\begin{aligned}
V_h^{\pi^*}(s_h) - V_h^{\hat{\pi}^k}(s_h) &\le \sum_{i=h}^{H} \mathbb{E}_{\pi^*}\left[\underbrace{\zeta_i^k(s_i,a_i)}_{\le 2b_i^k(s_i,a_i)}|s_h\right] - \sum_{i=h}^{H} \mathbb{E}_{\hat{\pi}^k}\left[\underbrace{\zeta_i^k(s_i,a_i)}_{\ge 0}|s_h\right] \\
&+ \sum_{i=h}^{H} \mathbb{E}_{\pi^*}\left[\underbrace{\langle \hat{Q}_i^k(s_i,\cdot), \pi_i^*(\cdot|s_i) - \hat{\pi}_i^k(\cdot|s_i)\rangle}_{\le 0}|s_h\right] \\
&\le 2\mathbb{E}_{\pi^*}\left[\sum_{i=h}^{H} b_i^k(s_i,a_i)|s_h\right]
\end{aligned}
$$

where the second inequality also follows from Lemma A.1. □

Now we prove that $b_h^k = \beta_k \cdot \|\phi_h(\cdot,\cdot)\|_{(\Sigma_h^k)^{-1}}$ is indeed a valid uncertainty quantifer function.

Lemma A.3. *There exists an absolute constant $c_1 > 0$ such that for any $\delta > 0$, if we choose $\beta_k = \beta_k(\delta) := c_1 \cdot dH \log(dHk/\delta)$ in Algorithm 1, then with probability at least $1-\delta$:*

$$
\forall(k,h,s,a) \in [K]\times[H]\times\mathcal{S}\times\mathcal{A}, |(T_h\hat{V}_{h+1}^k)(s,a) - (\hat{T}_h^k\hat{V}_{h+1}^k)(s,a)| \le \beta_k \cdot \|\phi_h(s,a)\|_{(\Sigma_h^k)^{-1}}.
$$

Proof. Let w_h^k such that $T_h\hat{V}_{h+1}^k = \langle \phi_h, w_h^k\rangle$ (such a w_h^k exists due to Lemma E.3). Recall that $\hat{T}_h^k\hat{V}_{h+1}^k = \langle \phi_h, \hat{w}_h^k\rangle$. We have

$$
\begin{aligned}
&(T_h\hat{V}_{h+1}^k)(s,a) - (\hat{T}_h^k\hat{V}_{h+1}^k)(s,a) = \phi_h(s,a)^T(w_h^k - \hat{w}_h^k) \\
&= \phi_h(s,a)^T w_h^k - \phi_h(s,a)^T(\Sigma_h^k)^{-1}\sum_{t=1}^{k-1}\phi_h(s_h^t,a_h^t)(r_h^t + \hat{V}_{h+1}^k(s_{h+1}^t)) \\
&= \underbrace{\phi_h(s,a)^T w_h^k - \phi_h(s,a)^T(\Sigma_h^k)^{-1}\sum_{t=1}^{k-1}\phi_h(s_h^t,a_h^t)\cdot(T_h\hat{V}_{h+1}^k)(s_h^t,a_h^t)}_{(i)} \\
&+ \underbrace{\phi_h(s,a)^T(\Sigma_h^k)^{-1}\sum_{t=1}^{k-1}\phi_h(s_h^t,a_h^t)\cdot\left[(T_h\hat{V}_{h+1}^k)(s_h^t,a_h^t) - r_h^t - \hat{V}_{k+1}^k(s_{h+1}^t)\right]}_{(ii)}.
\end{aligned}
$$

We bound term (i) by

$$
\begin{aligned}
|(i)| &= \left|\phi_h(s,a)^T w_h^k - \phi_h(s,a)^T(\Sigma_h^k)^{-1}\sum_{t=1}^{k-1}\phi_h(s_h^t,a_h^t)\phi_h(s_h^t,a_h^t)^T w_h^k\right| \\
&= \left|\lambda\phi_h(s,a)^T(\Sigma_h^k)^{-1}w_h^k\right| \le \lambda\cdot\|\phi_h(s,a)\|_{(\Sigma_h^k)^{-1}}\cdot\|w_h^k\|_{(\Sigma_h^k)^{-1}} \\
&\le \lambda\cdot\|\phi_h(s,a)\|_{(\Sigma_h^k)^{-1}}\cdot\|w_h^k\|_2\sqrt{\|(\Sigma_h^k)^{-1}\|} \\
&\le 2H\sqrt{d\lambda}\cdot\|\phi_h(s,a)\|_{(\Sigma_h^k)^{-1}}.
\end{aligned}
$$

Let $\eta_h^t = (T_h\hat{V}_{h+1}^t)(s_h^t,a_h^t) - r_h^t - \hat{V}_{h+1}^t(s_{h+1}^t)$. We have

$$
|(ii)| = \left|\phi_h(s,a)^T(\Sigma_h^k)^{-1}\sum_{t=1}^{k-1}\phi_h(s_h^t,a_h^t)\cdot\eta_t\right| \le \|\phi_h(s,a)\|_{(\Sigma_h^k)^{-1}}\cdot\underbrace{\|\sum_{t=1}^{k-1}\phi_h(s_h^t,a_h^t)\cdot\eta_h^t\|_{(\Sigma_h^k)^{-1}}}_{(iii)}.
$$

By Lemma E.4-E.5-E.12, for any $\delta > 0$, with probability at least $1-\delta$, we have:

$$(iii)^2 \le 4H^2 \left[\frac{d}{2}\log\left(\frac{k+\lambda}{\lambda}\right) + d\log(1+4H\sqrt{dk/\lambda}/\epsilon) + d^2\log(1+8\sqrt{d}\beta_k^2/(\lambda\epsilon^2)) + \log(1/\delta)\right] + \frac{8k^2\epsilon^2}{\lambda}.$$

Choosing $\epsilon = dH/k$ and $\lambda = 1$, combining terms (i), (ii), (iii) and using the union bound complete the proof. $\square$

Combing Lemma A.1 and Lemma F.1 via the union bound, we have the following main lemma for this step.

Lemma A.4. *There exists an absolute constant $c_1 > 0$ such that if we choose $\beta_k = \beta_k(\delta) := c_1 \cdot dH\log(dHK/\delta)$ in Algorithm 1, under Assumption 4.1, then with probability at least $1-2\delta$, we have:*

$$\forall (s_1, h, k) \in \mathcal{S}\times[H]\times[K], \mathrm{SubOpt}(\hat{\pi}^k; s_1) \le 2\beta_k(\delta)\mathbb{E}_{\pi^*}\left[\sum_{h=1}^{H}\|\phi_h(s_h,a_h)\|_{(\Sigma_h^k)^{-1}}|\mathcal{F}_{k-1}, s_1\right].$$

Step 2: Generalization. Next, we bound $\mathrm{SubOpt}(\hat{\pi})$ in terms of its empirical bootstraps $\mathrm{SubOpt}(\hat{\pi}^k; s_1^k)$ (plus a generalization error). Working with individual observed initial states s_1^k is helpful in constructing a complete trajectory $(s_1^k, a_1^k, \ldots, s_H^k, a_H^k)$ of μ which can then be connected to the trajectory of an optimal policy π^* via the optimal-policy concentrability assumption in the next step. We first state and prove the an online-to-batch argument which improves the generalization error of the original online-to-batch argument (Cesa-Bianchi et al., 2004) from $\mathcal{O}(\frac{1}{\sqrt{K}})$ to $\mathcal{O}(\frac{\log\log K}{K})$. This result could be of independent interest.

Lemma A.5 (Improved online-to-batch argument)**.** *Let $\{X_k\}$ be any real-valued stochastic process adapted to the filtration $\{\mathcal{F}_k\}$, i.e. X_k is $\mathcal{F}_k$-measurable. Suppose that for any k, $X_k \in [0,H]$ almost surely for some $H>0$. For any $K>0$, with probability at least $1-\delta$, we have:*

$$\sum_{k=1}^{K}\mathbb{E}\left[X_k|\mathcal{F}_{k-1}\right] \le 2\sum_{k=1}^{K}X_k + \frac{16}{3}H\log(\log_2(KH)/\delta) + 2.$$

Proof of Lemma A.5 . Let $Z_k = X_k - \mathbb{E}\left[X_k|\mathcal{F}_{k-1}\right]$ and $f(K) = \sum_{k=1}^{K}\mathbb{E}\left[X_k|\mathcal{F}_{k-1}\right]$. We have Z_k is a real-valued difference martingale with the corresponding filtration $\{\mathcal{F}_k\}$ and that

$$V := \sum_{k=1}^{K}\mathbb{E}\left[Z_k^2|\mathcal{F}_{k-1}\right] \le \sum_{k=1}^{K}\mathbb{E}\left[X_k^2|\mathcal{F}_{k-1}\right] \le H\sum_{k=1}^{K}\mathbb{E}\left[X_k|\mathcal{F}_{k-1}\right] = Hf(K).$$

Note that $|Z_k| \le H$ and $f(K) \in [0, KH]$ and let $m = \log_2(KH)$. Also note that $f(K) = \sum_{k=1}^{K}X_k - \sum_{k=1}^{K}Z_k \ge -\sum_{k=1}^{K}Z_k$. Thus if $\sum_{k=1}^{K}Z_k \le -1$, we have $f(K)\ge 1$. For any $t > 0$, leveraging the peeling technique (Bartlett et al., 2005), we have:

$$\mathbb{P}\left(\sum_{k=1}^{K}Z_k \le -\frac{2Ht}{3} - \sqrt{4Hf(K)t} - 1\right) = \mathbb{P}\left(\sum_{k=1}^{K}Z_k \le -\frac{2Ht}{3} - \sqrt{4Hf(K)t} - 1, f(K)\in[1,KH]\right)$$

$$
\begin{aligned}
&\leq \sum_{i=1}^{m} \mathbb{P}\left(\sum_{k=1}^{K} Z_k \leq -\frac{2Ht}{3} - \sqrt{4Hf(K)t} - 1, f(K) \in [2^{i-1}, 2^i)\right) \\
&\leq \sum_{i=1}^{m} \mathbb{P}\left(\sum_{k=1}^{K} Z_k \leq -\frac{2Ht}{3} - \sqrt{4H2^{i-1}t} - 1, V \leq H2^i, f(K) \in [2^{i-1}, 2^i)\right) \\
&\leq \sum_{i=1}^{m} \mathbb{P}\left(\sum_{k=1}^{K} Z_k \leq -\frac{2Ht}{3} - \sqrt{2H2^i t}, V \leq H2^i\right) \\
&\leq \sum_{i=1}^{m} e^{-t} = me^{-t},
\end{aligned}
$$

where the first equation is by that $\sum_{k=1}^{K} Z_k \leq -\frac{2Ht}{3} - \sqrt{4Hf(K)t} - 1 \leq -1$ thus $f(K) \geq 1$, the second inequality is by that $V \leq Hf(K)$, and the last inequality is by Lemma E.13. Thus, with probability at least $1 - me^{-t}$, we have:

$$
\sum_{k=1}^{K} X_k - f(K) = \sum_{k=1}^{K} Z_k \geq -\frac{2Ht}{3} - \sqrt{4Hf(K)t} - 1.
$$

The above inequality implies that $f(K) \leq 2\sum_{k=1}^{K} X_k + 4Ht/3 + 2 + 4Ht$, due to the simple inequality: if $x \leq a\sqrt{x} + b$, $x \leq a^2 + 2b$. Then setting $t = \log(m/\delta)$ completes the proof. $\square$

Now, we state and prove the main lemma for the generalization step.

Lemma A.6. *With probability at least* $1 - \delta$*, we have:*

$$
\sum_{k=1}^{K} \mathrm{SubOpt}(\hat{\pi}^k) \leq 2\sum_{k=1}^{K} \mathrm{SubOpt}(\hat{\pi}^k; s_1^k) + \frac{16}{3} H \log(\log_2(KH)/\delta) + 2.
$$

Remark 7. *Lemma A.6 is similar to (Nguyen-Tang et al., 2022, Lemma A.3) but here we obtain the* $\log(\log(K))$ *error while the latter only obtains the* $\sqrt{K}$ *error. The key for this improvement is to use the peeling technique and the variance information via Freedman inequality.*

Proof of Lemma A.6. Let $X_k = \mathrm{SubOpt}(\hat{\pi}^k; s_1^k)$ and recall that $\mathcal{F}_k = \sigma\left(\{(s_h^t, a_h^t, r_h^t)\}_{h\in[H]}^{t\in[k]}\right)$. As $\hat{\pi}^k$ is $\mathcal{F}_{k-1}$-measurable, and s_1^k is $\mathcal{F}_k$-measurable, we have that X_k is $\mathcal{F}_k$-measurable and $\mathbb{E}\left[X_k|\mathcal{F}_{k-1}\right] = \mathbb{E}\left[\mathrm{SubOpt}(\hat{\pi}^k; s_1^k)|\mathcal{F}_{k-1}\right] = \mathrm{SubOpt}(\hat{\pi}^k)$. Note that $X_k \in [0, H]$. Thus, by Lemma A.5, with probability at least $1 - \delta$, we have:

$$
\sum_{k=1}^{K} \mathbb{E}\left[X_k|\mathcal{F}_{k-1}\right] \leq 2\sum_{k=1}^{K} X_k + \frac{16}{3} H \log(\log_2(KH)/\delta) + 2.
$$

As $\hat{\pi}$ is uniformly sampled from $\{\hat{\pi}^k\}_{k\in[K]}$ conditioned on $\mathcal{D}$,

$$
K \cdot \mathbb{E}\left[\mathrm{SubOpt}(\hat{\pi})|\mathcal{D}\right] = \sum_{k=1}^{K} \mathrm{SubOpt}(\hat{\pi}^k) = \sum_{k=1}^{K} \mathbb{E}\left[X_k|\mathcal{F}_{k-1}\right].
$$

Thus we can complete the proof. $\square$

Step 3: Marginalized importance sampling. This step is the key in our proof to handle the distributional shift under the OPC. The high-level idea is to use importance sampling for the marginalized visitation density functions.

Lemma A.7. *Under Assumption 4.1, for any $h \in [H]$, with probability at least $1-\delta$, we have:*

$$\sum_{k=1}^{K} \mathbb{E}_{(s_h,a_h)\sim d_h^*(\cdot,\cdot)} \left[\|\phi_h(s_h,a_h)\|_{(\Sigma_h^k)^{-1}} \middle| \mathcal{F}_{k-1}, s_1^k \right] \leq \sum_{k=1}^{K} \frac{d_h^*(s_h^k,a_h^k)}{d_h^\mu(s_h^k,a_h^k)} \|\phi_h(s_h^k,a_h^k)\|_{(\Sigma_h^k)^{-1}} + \sqrt{\frac{1}{\lambda}\log(1/\delta)} \sqrt{\sum_{k=1}^{K} \left(\frac{d_h^*(s_h^k,a_h^k)}{d_h^\mu(s_h^k,a_h^k)} \right)^2}.$$

Proof of Lemma A.7. Let $Z_h^k := \frac{d_h^*(s_h^k,a_h^k)}{d_h^\mu(s_h^k,a_h^k)} \|\phi_h(s_h^k,a_h^k)\|_{(\Sigma_h^k)^{-1}}$. We have Z_h^k is $\mathcal{F}_k$-measurable, and by Assumption 4.1, we have,

$$|Z_h^k| \leq \frac{d_h^*(s_h^k,a_h^k)}{d_h^\mu(s_h^k,a_h^k)} \|\phi_h(s_h^k,a_h^k)\|_2 \sqrt{\|(\Sigma_h^k)^{-1}\|} \leq 1/\sqrt{\lambda} \frac{d_h^*(s_h,a_h)}{d_h^\mu(s_h^k,a_h^k)} < \infty,$$

and

$$\mathbb{E}\left[Z_h^k | \mathcal{F}_{k-1}, s_1^k \right] = \mathbb{E}_{(s_h,a_h)\sim d_h^\mu(\cdot,\cdot)} \left[\frac{d_h^*(s_h,a_h)}{d_h^\mu(s_h,a_h)} \|\phi_h(s_h,a_h)\|_{(\Sigma_h^k)^{-1}} \middle| \mathcal{F}_{k-1}, s_1^k \right].$$

Thus, by Lemma E.10, for any $h \in [H]$, with probability at least $1-\delta$, we have:

$$\begin{aligned} &\sum_{k=1}^{K} \mathbb{E}_{(s_h,a_h)\sim d_h^*(\cdot,\cdot)} \left[\|\phi_h(s_h,a_h)\|_{(\Sigma_h^k)^{-1}} \middle| \mathcal{F}_{k-1}, s_1^k \right] \\ &= \sum_{k=1}^{K} \mathbb{E}_{(s_h,a_h)\sim d_h^\mu(\cdot,\cdot)} \left[\frac{d_h^*(s_h,a_h)}{d_h^\mu(s_h,a_h)} \|\phi_h(s_h,a_h)\|_{(\Sigma_h^k)^{-1}} \middle| \mathcal{F}_{k-1}, s_1^k \right] \\ &\leq \sum_{k=1}^{K} \frac{d_h^*(s_h^k,a_h^k)}{d_h^\mu(s_h^k,a_h^k)} \|\phi_h(s_h^k,a_h^k)\|_{(\Sigma_h^k)^{-1}} + \sqrt{\frac{1}{\lambda}\log(1/\delta)} \sqrt{\sum_{k=1}^{K} \left(\frac{d_h^*(s_h^k,a_h^k)}{d_h^\mu(s_h^k,a_h^k)} \right)^2}. \end{aligned} \tag{3}$$

where the second equation is valid due to Assumption 4.1.

□

Theorem 1 is the direct combination of Lemma A.4-A.6-A.7 via the union bound.

A.2 Proof for the last-iteration policy in Theorem 1

In this part, we provide the proof for the bound of $\hat{\pi}_{PEVI}$ in Theorem 1. The proof is similar to that of $\hat{\pi}_{unif}$ except that we directly reason on the elliptical potential $\beta(\delta)\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\|\phi_h(s_h,a_h)\|_{(\Sigma_h^k)^{-1}} | s_1^k, \mathcal{F}_{k-1} \right]$ rather than the empirical sub-optimality metric $\text{SubOpt}(\hat{\pi}^k; s_1^k)$. We establish the proof steps in the following.

Step 1: The sub-optimality bound of the last-iteration policy $\hat{\pi}^{last}$. The first step directly follows the original proof of PEVI (Jin et al., 2021) with a simple modification. In particular, with probability at least $1-2\delta$, we have:

$$\text{SubOpt}(\hat{\pi}^{last}) \leq \min\left\{ H, 2\beta(\delta) \mathbb{E}_{s_1\sim d_1} \left[\sum_{h=1}^{H} \mathbb{E}_{\pi^*}\left[\|\phi_h(s_h,a_h)\|_{\Sigma_h^{-1}} | s_1 \right] \right] \right\},$$

where $\mathbb{E}_{\pi^*}$ is with respect to the trajectory $(s_1, a_1, \ldots, s_H, a_H)$ induced by π^*.

Step 2: Generalization. The key idea for this step is to bound the expectation $\mathbb{E}_{s_1 \sim d_1}[\cdot|s_1]$ in terms of the empirical quantities $[\cdot|s_1^k]$ (plus a generalization error). The empirical quantities $[\cdot|s_1^k]$ is useful in constructing a complete trajectory $(s_1^k, a_1^k, \dots, s_H^k, a_H^k)$ of μ which can then be connected to the trajectory of the optimal policy π^* via the density dominance assumption. However, as Σ_h depends on $\{s_1^k\}_{k \in [K]}$, the generalization error from the expectation $\mathbb{E}_{s_1 \sim d_1}[\cdot|s_1]$ to the empirical quantities $[\cdot|s_1^k]$ cannot guarantee. [4] Instead we use a simple trick to break such data dependence and form a valid martingale for strong generalization. Let $Z_h^k := 2\beta(\delta) \sum_{h=1}^H \mathbb{E}_{\pi^*} \left[\|\phi_h(s_h, a_h)\|_{(\Sigma_h^k)^{-1}} | s_1^k, \mathcal{F}_{k-1} \right]$. We have:

$$
\begin{aligned}
&K \cdot \mathrm{SubOpt}(\hat{\pi}^{last}) \leq \sum_{k=1}^K \min \left\{ H, 2\beta(\delta) \mathbb{E}_{s_1 \sim d_1} \left[\sum_{h=1}^H \mathbb{E}_{\pi^*} \left[\|\phi_h(s_h, a_h)\|_{\Sigma_h^{-1}} \Big| s_1, \mathcal{F}_K \right] \right] \right\} \\
&\leq \sum_{k=1}^K \min \left\{ H, 2\beta(\delta) \mathbb{E}_{s_1 \sim d_1} \left[\sum_{h=1}^H \mathbb{E}_{\pi^*} \left[\|\phi_h(s_h, a_h)\|_{(\Sigma_h^k)^{-1}} \Big| s_1, \mathcal{F}_{k-1} \right] \right] \right\} \\
&= \sum_{k=1}^K \min \left\{ H, \mathbb{E}\left[Z_k | \mathcal{F}_{k-1} \right] \right\} \\
&\leq \sum_{k=1}^K \mathbb{E}\left[\min\{H, Z_k\} | \mathcal{F}_{k-1} \right] \\
&\leq 2 \sum_{k=1}^K \min\{H, Z_k\} + \frac{16}{3} H \log(\log_2(KH)/\delta) + 2 \\
&= 2 \sum_{k=1}^K \min \left\{ H, 2\beta(\delta) \sum_{h=1}^H \mathbb{E}_{\pi^*} \left[\|\phi_h(s_h, a_h)\|_{(\Sigma_h^k)^{-1}} | s_1^k, \mathcal{F}_{k-1} \right] \right\} + \frac{16}{3} H \log(\log_2(KH)/\delta) + 2,
\end{aligned}
$$

where the second inequality is by $\min\{a, b\} \leq \min\{a, c\}$ if $b \leq c$ and $\Sigma_h^{-1} \preceq (\Sigma_h^k)^{-1}, \forall k \in [K+1]$, the third inequality is by Jensen's inequality (as $f(x) = \min\{H, x\}$ is convex), and the fourth inequality is by Lemma A.5 (where the use of $(\Sigma_h^k)^{-1}$ in the place of Σ_h^{-1} is crucial to form a valid martingale for applying Lemma A.5 and $\min\{H, Z_k\} \leq H$).

Step 3: Marginalized importance sampling. This step is the same as the marginalized importance sampling step for $\hat{\pi}^{mix}$.

A.3 Proof of Corollary 1

We give a proof for Corollary 1. Using Theorem 1, it suffices to prove the following lemma.

Lemma A.8. *Under Assumption 4.1-4.2, for any $h \in [H]$, we have:*

$$
\sum_{k=1}^K \frac{d_h^*(s_h^k, a_h^k)}{d_h^\mu(s_h^k, a_h^k)} \|\phi_h(s_h^k, a_h^k)\|_{(\Sigma_h^k)^{-1}} \leq \iota_h^{-1} \sqrt{2Kd \log(1 + K/d)}.
$$

Proof of Lemma A.8. We have:

$$
\begin{aligned}
\sum_{k=1}^K \frac{d_h^*(s_h^k, a_h^k)}{d_h^\mu(s_h^k, a_h^k)} \|\phi_h(s_h^k, a_h^k)\|_{(\Sigma_h^k)^{-1}} &\leq \iota_h^{-1} \sum_{k=1}^K \|\phi_h(s_h^k, a_h^k)\|_{(\Sigma_h^k)^{-1}} \leq \iota_h^{-1} \sqrt{K \sum_{k=1}^K \|\phi_h(s_h^k, a_h^k)\|^2_{(\Sigma_h^k)^{-1}}} \\
&\leq \iota_h^{-1} \sqrt{2K \log \frac{\det \Sigma_h^{K+1}}{\det(I)}} \leq \iota_h^{-1} \sqrt{2K} \sqrt{d \log(1 + K/d)}.
\end{aligned}
$$

[4] Or at least we must use the uniform convergence argument to overcome this data-dependent structure which makes up a large and unnecessary generalization error.

where the first inequality is by Assumption 4.1, the second inequality is by Cauchy-Schwartz inequality, and the last two inequalities are by (Abbasi-yadkori et al., 2011, Lemma 11). $\square$

We also provide an information-theoretical lower bound for Assumption 4.1-4.2.

Theorem 5. *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 exist a linear MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r, d_0)$ 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:*

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

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

A.4 Data-dependent bounds under arbitrary data coverage

Note that Theorem 1 requires the OPC assumption (Assumption 4.1) to be valid. In practice, as we do not have control over the behavior policy, it happens that the behavior policy does not fully cover all the trajectories of the optimal policy, thus the OPC assumption might fail to hold. This raises the question of how much an offline algorithm can suffer when it learns from the offline data of arbitrary coverage.

To tackle this issue, we construct a new MDP $\bar{\mathcal{M}}$ under which the trajectories of the optimal policy $\bar{\pi}^*$ (with respect to $\bar{\mathcal{M}}$) are best covered by the behavior policy μ. Then, the sub-optimality gap incurred by the under-coverage data must be the gap between the original optimal policy π^* and the data-supported optimal policies $\bar{\pi}^*$.

Augmented MDP. For any small positive $\bar{\epsilon} > 0$, we consider the augmented MDP $\bar{\mathcal{M}} = (\mathcal{S} \cup \{\bar{s}_{h+1}\}_{h\in[H]}, \mathcal{A}, \bar{\mathbb{P}}, \bar{r}, H, d_1)$, where for any $h \in [H]$

$$\bar{\mathbb{P}}_h(\cdot|s_h, a_h) = \begin{cases} \mathbb{P}_h(\cdot|s_h, a_h) & \text{if } (s_h, a_h) \in \mathcal{C}_h^\mu \\ \delta_{\bar{s}_{h+1}} & \text{if } (s_h, a_h) \notin \mathcal{C}_h^\mu \end{cases}, \bar{r}_h(s_h, a_h) = \begin{cases} r_h(s_h, a_h) & \text{if } (s_h, a_h) \in \mathcal{C}_h^\mu \\ -\bar{\epsilon}/H & \text{if } (s_h, a_h) \notin \mathcal{C}_h^\mu \end{cases}$$

Here $\bar{\mathcal{M}}$ extends the original state $\mathcal{S}$ to include arbitrary states $\{\bar{s}_{h+1}\}_{h\in[H]}$ where $\bar{s}_{h+1} \notin \mathcal{S}$. The transition distributions and the reward functions are the same as the original $\mathcal{M}$ except at the infeasible state-action $(s_h, a_h) \notin \mathcal{C}_h^\mu$ where the augmented MDP always absorbs into the dummy state $\bar{s}_{h+1}$ and yields small negative reward. Under $\bar{\mathcal{M}}$, we denote the corresponding marginal state-visitation density by $\bar{d}^\pi$ and the corresponding optimal policy $\bar{\pi}^*$. We abbreviate $\bar{d}^{\bar{\pi}^*} = \bar{d}^*$. Our augmented MDP construction is similar to the construction in Yin & Wang (2021) except that we allow an arbitrary small negative reward $-\bar{\epsilon}/H$ in unsupported state-action pairs. This design guarantees that the optimal policy under $\bar{\mathcal{M}}$ is dominated by μ (Lemma A.9). The following theorem (Theorem 6) provides a generic (instance-agnostic) bound that works under arbitrary data coverage.

Theorem 6. *Under Assumption 4.1, w.p.a.l. $1 - 4\delta$ over the randomness of $\mathcal{D}$, we have:*

$$\mathrm{SubOpt}(\hat{\pi}^{mix}) \vee \mathrm{SubOpt}(\hat{\pi}^{last}) \leq gap_{support} + \frac{4\beta(\delta)}{K}\sum_{h=1}^H\sum_{k=1}^K \frac{\bar{d}_h^*(s_h^k, a_h^k)}{d_h^\mu(s_h^k, a_h^k)}\|\phi_h(s_h^k, a_h^k)\|_{(\Sigma_h^k)^{-1}}$$

$$+ \frac{4\beta(\delta)}{K}\sum_{h=1}^H\sqrt{\log\left(\frac{H}{\delta}\right)\sum_{k=1}^K\left(\frac{\bar{d}_h^*(s_h^k, a_h^k)}{d_h^\mu(s_h^k, a_h^k)}\right)^2} + \frac{2}{K} + \frac{16H}{3K}\log\left(\frac{\log_2(KH)}{\delta}\right),$$

where $gap_{support} := \sum_{h=1}^{H} \int_{(\mathcal{C}_h^{\pi})^c} d_h^{\pi}(s_h, a_h) ds_h a_h + \bar{\epsilon}$ is the sub-optimality gap incurred in locations that are supported by π^* but unsupported by μ, and $\bar{d}^*$ is the marginal visitation density of the optimal policy under the augmented MDP $\bar{\mathcal{M}}$.

Theorem 6 is valid for any data coverage. When the behavior policy μ does not fully support the trajectories of any optimal policy π^*, the algorithm must suffer a constant sub-optimality gap $\text{gap}_{\text{support}}$ which is incurred by the total rewards at the locations supported by the optimal policy but not by the behavior policy. When the OPC assumption (Assumption 4.1) holds, $\text{gap}_{\text{support}} = 0$ and $\bar{d}_h^* = d_h^*$ (Lemma A.9), and Theorem 6 reduces into Theorem 1.

Before proving Theorem 6, we provide and prove useful properties of the augmented MDP $\bar{\mathcal{M}}$ from the perspective of the original MDP $\mathcal{M}$.

Lemma A.9. *Consider any $(h, s_h, a_h) \in [H] \times \bar{\mathcal{S}} \times \mathcal{A}$, and any policy π.*

(a) $\bar{d}_h^{\mu}(s_h) = d_h^{\mu}(s_h), \forall (h, s_h, a_h) \in [H] \times \bar{\mathcal{S}} \times \mathcal{A}$.

(b) For any policy π, if $(s_h, a_h) \notin \mathcal{C}_h^{\mu}$, $\bar{Q}_h^{\pi}(s_h, a_h) = -(H - h + 1)$.

(c) For any $(h, s_h, a_h) \in [H] \times \bar{\mathcal{S}} \times \mathcal{A}$, if $\bar{d}_h^*(s_h, a_h) > 0$, then $\bar{d}_h^{\mu}(s_h, a_h) > 0$.

(d) Under Assumption 4.1, $\bar{d}_h^* = d_h^*, \forall h \in [H]$.

Proof. **Part** (a)**.** We prove (a) by induction. First, we have:

$$\forall s_1 \in \bar{\mathcal{S}}, \bar{d}_1^{\pi}(s_1) = \bar{d}_1(s_1) = d_1(s_1) = d_1^{\pi}(s_1).$$

Suppose that for some $h \in [H]$, $\bar{d}_h^{\pi}(s_h) = d_h^{\pi}(s_h), \forall s_h \in \mathcal{C}_h^{\mu}$. Consider any $s_{h+1} \in \bar{\mathcal{S}}$. We have:

$$\begin{aligned}
\bar{d}_{h+1}^{\mu}(s_{h+1}) &= \int \bar{p}_h(s_{h+1}|s_h, a_h) \bar{d}_h^{\mu}(s_h, a_h) d(s_h a_h) \\
&= \int_{\mathcal{C}_h^{\mu}} \bar{p}_h(s_{h+1}|s_h, a_h) \bar{d}_h^{\mu}(s_h, a_h) d(s_h a_h) + \int_{\bar{\mathcal{C}}_h^{\mu}} \bar{p}_h(s_{h+1}|s_h, a_h) \bar{d}_h^{\mu}(s_h, a_h) d(s_h a_h) \\
&= \int_{\mathcal{C}_h^{\mu}} p_h(s_{h+1}|s_h, a_h) d_h^{\mu}(s_h, a_h) d(s_h a_h) + \int_{\bar{\mathcal{C}}_h^{\mu}} \bar{p}_h(s_{h+1}|s_h, a_h) \underbrace{d_h^{\mu}(s_h, a_h)}_{=0} d(s_h a_h) \\
&= \int_{\mathcal{C}_h^{\mu}} p_h(s_{h+1}|s_h, a_h) d_h^{\mu}(s_h, a_h) d(s_h a_h) + \int_{\bar{\mathcal{C}}_h^{\mu}} p_h(s_{h+1}|s_h, a_h) \underbrace{d_h^{\mu}(s_h, a_h)}_{=0} d(s_h a_h) \\
&= d_{h+1}^{\mu}(s_{h+1}),
\end{aligned}$$

where the third equation is due to that $\bar{p}_h(s_{h+1}|s_h, a_h) = p_h(s_{h+1}|s_h, a_h)$ for $(s_h, a_h) \in \mathcal{C}_h^{\mu}$, $\bar{d}_h^{\mu}(s_h, a_h) = d_h^{\mu}(s_h, a_h)$ for any (s_h, a_h) (by the induction step), and the fourth equation is by $d_h^{\mu}(s_h, a_h) = 0$ for any $(s_h, a_h) \notin \mathcal{C}_h^{\mu}$ (by definition).

Part (b)**.** For any $(s_h, a_h) \notin \mathcal{C}_h^{\mu}$, the feasible trajectory must admit the following form: $(s_h, a_h, \bar{s}_{h+1}, a_{h+1}, \ldots, \bar{s}_H, a_H, \bar{s}_{H+1})$ which has zero cumulative reward under $\bar{r}$ by definition.

Part (c)**.** By (a), we now replace $\bar{d}_h^{\mu}$ by d_h^{μ} in (c) and prove (c) by induction. At initial state $h = 1$, if $\bar{d}_1^*(s_1, a_1) = d_1(s_1)\bar{\pi}_1^*(a_1|s_1) > 0$, we have $d_1(s_1) > 0$ and $\bar{\pi}_1^*(a_1|s_1) > 0$. Thus a_1 must be an optimal action given s_1 (under $\bar{\mathcal{M}}$). Suppose that $(s_1, a_1) \notin \mathcal{C}_1^{\mu}$. By (b), $\bar{Q}_1^*(s_1, a_1) = -H$. Let any $\tilde{a}_1$ such that $(s_1, \tilde{a}_1) \in \mathcal{C}_1^{\mu}$ (such $\tilde{a}_1$ exists as $s_1 \in \mathcal{C}_1^{\mu}$). Then, we have $\bar{Q}^*(s_1, \tilde{a}_1) = r_1(s_1, \tilde{a}_1) + \mathbb{E}_{\bar{\pi}^*}\left[\sum_{i=2}^{H} \bar{r}_i \middle| s_1, \tilde{a}_1\right] \geq -(H-1) > \bar{Q}_1^*(s_1, a_1)$. This contradicts that a_1 must be an optimal action given s_1 (under $\bar{\mathcal{M}}$).

Now assume that we have (c) for some $h \geq 1$. We will prove (c) for $h+1$. Indeed, consider any (s_{h+1}, a_{h+1}) such that $\bar{d}^*_{h+1}(s_{h+1}, a_{h+1}) > 0$. Then we must have $s_{h+1} \in C^*_{h+1}$ and a_{h+1} is an optimal action given s_{h+1} (under $\bar{\mathcal{M}}$). Since $s_{h+1} \in C^*_{h+1}$, there must be some $(s_h, a_h) \in \mathcal{C}^*_h$ such that $\bar{p}_h(s_{h+1}|s_h, a_h) > 0$. By induction, we have $(s_h, a_h) \in \mathcal{C}^*_h \subseteq \mathcal{C}^\mu_h$. Thus, $0 < \bar{p}_h(s_{h+1}|s_h, a_h) = p_h(s_{h+1}|s_h, a_h)$. Hence, $s_{h+1} \in \mathcal{C}^\mu_{h+1}$. Given $s_{h+1} \in \mathcal{C}^\mu_{h+1}$ and a_{h+1} is an optimal action given s_{h+1} (under $\bar{\mathcal{M}}$), similar to the base case, we must have $(s_{h+1}, a_{h+1}) \in \mathcal{C}^\mu_{h+1}$.

Part (d)**.** Under Assumption 4.1, we can prove by induction from $H, H-1, \ldots, 1$ that: $\bar{Q}^*_h(s_h, a_h) = Q^*_h(s_h, a_h)$ and $\bar{V}^*_h(s_h) > \bar{Q}^*_h(s_h, \tilde{a}_h)$ if $d^*_h(s_h) > 0$ and $\pi^*_h(a_h|s_h) > 0$ and $\pi^*_h(\tilde{a}_h|s_h) = 0$. This then implies (d). □

Now we are ready to prove the result in this subsection.

Proof of Theorem 6. For any policy π, we have:

$$
\begin{aligned}
\mathbb{E}_{\pi,\mathcal{M}}[r_h] - \mathbb{E}_{\pi,\bar{\mathcal{M}}}[\bar{r}_h] &= \int_{\mathcal{C}^\pi_h} r_h(s_h, a_h) d^\pi_h(s_h, a_h) ds_h a_h + \int_{(\mathcal{C}^\pi_h)^c} r_h(s_h, a_h) d^\pi_h(s_h, a_h) ds_h a_h \\
&\quad - \int_{\mathcal{C}^\pi_h} \bar{r}_h(s_h, a_h) \bar{d}^\pi_h(s_h, a_h) ds_h a_h - \int_{(\mathcal{C}^\pi_h)^c} \bar{r}_h(s_h, a_h) \bar{d}^\pi_h(s_h, a_h) ds_h a_h \\
&= \int_{(\mathcal{C}^\pi_h)^c} r_h(s_h, a_h) d^\pi_h(s_h, a_h) ds_h a_h + \epsilon_0 \in (0, \int_{(\mathcal{C}^\pi_h)^c} d^\pi_h(s_h, a_h) ds_h a_h + \bar{\epsilon}/H]
\end{aligned}
$$

where $(\mathcal{C}^\pi_h)^c$ denotes the complement of $\mathcal{C}^\pi_h$. Thus, we have:

$$
v^{\mathcal{M},\pi} - v^{\bar{\mathcal{M}},\pi} = \sum_{h=1}^H \mathbb{E}_{\pi,\mathcal{M}}[r_h] - \mathbb{E}_{\pi,\bar{\mathcal{M}}}[\bar{r}_h] \in (0, \sum_{h=1}^H \int_{(\mathcal{C}^\pi_h)^c} d^\pi_h(s_h, a_h) ds_h a_h + \bar{\epsilon}]. \tag{4}
$$

Define $gap_{support} := \sum_{h=1}^H \int_{(\mathcal{C}^\pi_h)^c} d^\pi_h(s_h, a_h) ds_h a_h + \bar{\epsilon}$, we have

$$
\begin{aligned}
& v^{\mathcal{M},\pi^*} - v^{\mathcal{M},\pi} = (v^{\bar{\mathcal{M}},\pi^*} - v^{\mathcal{M},\pi}) + (v^{\mathcal{M},\pi^*} - v^{\bar{\mathcal{M}},\pi^*}) \\
& \leq (v^{\bar{\mathcal{M}},\pi^*} - v^{\bar{\mathcal{M}},\pi}) + gap_{support} \\
& \leq (v^{\bar{\mathcal{M}},\bar{\pi}^*} - v^{\bar{\mathcal{M}},\pi}) + gap_{support} \\
& \leq \frac{4\beta(\delta)}{K} \sum_{h=1}^H \sum_{k=1}^K \frac{\bar{d}^*_h(s^k_h, a^k_h)}{\bar{d}^\mu_h(s^k_h, a^k_h)} \|\phi_h(s^k_h, a^k_h)\|_{(\Sigma^k_h)^{-1}} + \frac{4\beta(\delta)}{K} \sum_{h=1}^H \sqrt{\log\left(\frac{H}{\delta}\right) \sum_{k=1}^K \left(\frac{\bar{d}^*_h(s^k_h, a^k_h)}{\bar{d}^\mu_h(s^k_h, a^k_h)}\right)^2} \\
& + \frac{2}{K} + \frac{16H}{3K} \log\left(\frac{\log_2(KH)}{\delta}\right) + gap_{support} \\
& = \frac{4\beta(\delta)}{K} \sum_{h=1}^H \sum_{k=1}^K \frac{\bar{d}^*_h(s^k_h, a^k_h)}{d^\mu_h(s^k_h, a^k_h)} \|\phi_h(s^k_h, a^k_h)\|_{(\Sigma^k_h)^{-1}} + \frac{4\beta(\delta)}{K} \sum_{h=1}^H \sqrt{\log\left(\frac{H}{\delta}\right) \sum_{k=1}^K \left(\frac{\bar{d}^*_h(s^k_h, a^k_h)}{d^\mu_h(s^k_h, a^k_h)}\right)^2} \\
& + \frac{2}{K} + \frac{16H}{3K} \log\left(\frac{\log_2(KH)}{\delta}\right) + gap_{support},
\end{aligned}
$$

where the first inequality is by Eq. (4), the second inequality is by $\bar{\pi}^*$ is an optimal policy under $\bar{\mathcal{M}}$, the third inequality is by Theorem 1 (as under $\bar{\mathcal{M}}$, the single concentrability holds), and the last equality is by $\bar{d}^\mu = d^\mu$ (Lemma A.9). □

A.5 The Optimal-Policy Concentrability is necessary for learnability from offline data

This result affirms the necessity of the single-policy concentrability for learnability in offline RL.

Lemma A.10. *For any offline algorithm Algo$(\cdot)$, there exist an MDP instance $\mathcal{M}$ and an offline dataset $\mathcal{D}$ such that $\mathcal{M}$ generates $\mathcal{D}$, the single concentrability (Assumption 4.1) is not met, and Algo$(\mathcal{D})$ incurs a constant sub-optimality almost surely over the randomness of $\mathcal{D}$.*

Proof. Consider a class of bandit instances parameterized by $B(q_1, q_2, q_3)$ where $\{a_1, a_2, a_3\}$ is the shared action space of the class and q_i is the corresponding deterministic reward of action a_i for any $i \in [3]$. Now consider two bandit instances within the above class, namely $B_1 := B(0,0,1)$ and $B_2 := B(0,1,0)$, and construct dataset $\mathcal{D} = \{(b_t, r_t)\}_{t\in[K]}$ where $b_t = a_1, \forall t \in [K]$ and $r_t = 0, \forall t \in [K]$. As $\mathcal{D}$ only selects action a_1 and receives reward $r_t = 0$ while the reward of a_1 under both B_1 and B_2 is also 0, $\mathcal{D}$ is consistent with both B_1 and B_2 (in the sense that $\mathcal{D}$ could have been generated under either B_1 or B_2). Note that $\mathcal{D}$ does not satisfy the single concentrability (Assumption 4.1) under both B_1 and B_2 as $\mathcal{D}$ covers only action a_1 while the optimal actions for B_1 and B_2 are a_2 and a_3, respectively.

As the dataset $\mathcal{D}$ provides no information about a_2 and a_3, for any algorithm $\pi = \text{Algo}(\mathcal{D}) = (\pi(a_1), \pi(a_2), \pi(a_3)) \in \{(p_1, p_2, p_3) : p_i \geq 0, p_1 + p_2 + p_3 = 1\}$, $\pi(a_2)$ and $\pi(a_3)$ do not depend on $\mathcal{D}$. Without loss of generality, suppose $\pi(a_2) \geq \pi(a_3)$. Then, $\pi(a_3) \leq \frac{\pi(a_2)+\pi(a_3)}{2} \leq \frac{\pi(a_2)+\pi(a_3)+\pi(a_1)}{2} = \frac{1}{2}$. Thus, $\pi(a_1) + \pi(a_2) = 1 - \pi(a_3) \geq \frac{1}{2}$. Therefore, we have:

$$\text{SubOpt}(\text{Algo}(\mathcal{D}); B_2) = \pi(a_1) + \pi(a_2) \geq \frac{1}{2}.$$

It is crucial to note that the above inequality holds almost surely over the randomness of $\mathcal{D}$ as $\pi(a_2)$ and $\pi(a_3)$ are agnostic to $\mathcal{D}$. Thus, any Algo$(\mathcal{D})$ almost surely suffers a sub-optimality at least as large as $\frac{1}{2}$ under at least B_1 or B_2. $\square$

B Proof of Theorem 2

In this section, we provide the detailed proof for Theorem 2. We first state and prove a series of intermediate lemmas. The following lemma decomposes the sub-optimality of any policy into the gap information.

Lemma B.1 (Sub-optimality decomposition)**.** *We have:*

$$\forall (s_1, k) \in \mathcal{S} \times [K], V_1^*(s_1) - V_1^{\hat{\pi}^k}(s_1) = \mathbb{E}_{\hat{\pi}^k}\left[\sum_{h=1}^{H} \Delta_h(s_h, a_h) \middle| \mathcal{F}_{k-1}, s_1\right].$$

Proof of Lemma B.1. Conditioned on $\mathcal{F}_{k-1}$ and s_1, we have:

$$\begin{aligned} V_1^*(s_1) - V_1^{\hat{\pi}^k}(s_1) &= V_1^*(s_1) - Q_1^*(s_1, \hat{\pi}_1^k(s_1)) + Q_1^*(s_1, \hat{\pi}_1^k(s_1)) - Q_1^{\hat{\pi}^k}(s_1, \hat{\pi}_1^k(s_1)) \\ &= \mathbb{E}_{\hat{\pi}^k}\left[\Delta_1(s_1, a_1) | \mathcal{F}_{k-1}, s_1\right] + \mathbb{E}_{\hat{\pi}^k}\left[V_2^*(s_2) - Q_2^{\hat{\pi}^k}(s_2, a_2) | \mathcal{F}_{k-1}, s_1\right] \\ &= \mathbb{E}_{\hat{\pi}^k}\left[\Delta_1(s_1, a_1) | \mathcal{F}_{k-1}, s_1\right] + \mathbb{E}_{\hat{\pi}^k}\left[V_2^*(s_2) - V_2^{\hat{\pi}^k}(s_2) | \mathcal{F}_{k-1}, s_1\right]. \end{aligned}$$

Recursively applying the above equation over $h \in [H]$ and using the telescoping sum complete the proof. $\square$

The next lemma shows that any policy in the μ-supported policy class $\Pi(\mu)$ induce marginalized density functions that concentrate only within the support of the marginalized density functions under μ.

Lemma B.2 (Concentrability for the μ-supported policy class)**.** *For any $(\pi, h, s_h, a_h) \in \Pi(\mu) \times [H] \times \mathcal{S} \times \mathcal{A}$, we have:*

$$\frac{d_h^\pi(s_h, a_h)}{d_h^\mu(s_h, a_h)} < \infty.$$

Proof of Lemma B.2. Consider any $\pi \in \Pi(\mu)$. Note that the lemma statement is equivalent to

$$\forall h \in [H], \mathcal{S}_h^\pi \subseteq \mathcal{S}_h^\mu, \text{ and } \mathcal{SA}_h^\pi \subseteq \mathcal{SA}_h^\mu. \tag{5}$$

We prove Eq. (5) by induction with h. We have $\mathcal{S}_1^\pi = \mathcal{S}_1^\mu = \mathcal{S}_1$ by definition. For any $(s_1, a_1) \in \mathcal{SA}_1^\pi$, we have $s_1 \in \mathcal{S}_1$ and $\pi_1(a_1|s_1) > 0$. By the definition of $\Pi_1(\mu)$, $\mu_1(a_1|s_1) > 0$. Thus, we have $\mathcal{SA}_1^\pi \subseteq \mathcal{SA}_1^\mu$, i.e. Eq. (5) holds for $h = 1$.

Now assume that Eq. (5) holds for $h \geq 1$, we prove that Eq. (5) holds for $h+1$. Indeed, since $\mathcal{SA}_h^\pi \subseteq \mathcal{SA}_h^\mu$, we have:

$$\begin{aligned}
\mathcal{S}_{h+1}^\pi &= \{s_{h+1} \in \mathcal{S}_{h+1} : \exists (s_h, a_h) \in \mathcal{SA}_h^\pi \text{ such that } p_h(s_{h+1}|s_h, a_h) > 0\} \\
&\subseteq \{s_{h+1} \in \mathcal{S}_{h+1} : \exists (s_h, a_h) \in \mathcal{SA}_h^\mu \text{ such that } p_h(s_{h+1}|s_h, a_h) > 0\} = \mathcal{S}_{h+1}^\mu.
\end{aligned}$$

Now consider any $(s_{h+1}, a_{h+1}) \in \mathcal{SA}_{h+1}^\pi$. Then, we have $s_{h+1} \in \mathcal{S}_{h+1}^\pi \subseteq \mathcal{S}_{h+1}^\mu$ and $\pi_{h+1}(a_{h+1}|s_{h+1}) > 0$. By the definition of $\Pi_{h+1}(\mu)$, we have $\mu_h(a_{h+1}|s_{h+1}) > 0$. Thus, $(s_{h+1}, a_{h+1}) \in \mathcal{SA}_{h+1}^\mu$. □

The next lemma uses marginalized importance sampling to handle the distributional shift of the offline data to connect the sub-optimality of each $\hat{\pi}^k$ to the sub-optimality gap $\Delta_h(s_h^k, \hat{\pi}_h^k(s_h^k))$ under the behavior policy μ.

Lemma B.3 (Marginalized importance sampling for $\hat{\pi}^k$)**.** *Under Assumption 4.2, w.p.a.l. $1 - \delta$, we have For any $k \in [K]$, we have:*

$$\sum_{k=1}^K \text{SubOpt}(\hat{\pi}^k; s_1^k) \leq 2 \sum_{h=1}^H \kappa_h \sum_{k=1}^K \Delta_h(s_h^k, \hat{\pi}_h^k(s_h^k)) + \frac{16}{3} \kappa \log \log_2(K\kappa/\delta) + 2,$$

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

Proof of Lemma B.3. The key for the proof is to use the marginalized importance sampling due to Lemma B.3 and then apply the improved online-to-batch argument Lemma A.5. In particular, we have:

$$\begin{aligned}
\sum_{k=1}^K \text{SubOpt}(\hat{\pi}^k; s_1^k) &= \sum_{k=1}^K \mathbb{E}_{\hat{\pi}^k} \left[\sum_{h=1}^H \Delta_h(s_h, \hat{\pi}_h^k(s_h)) \middle| \mathcal{F}_{k-1}, s_1^k \right] \\
&\leq \sum_{k=1}^K \mathbb{E}_\mu \left[\sum_{h=1}^H \kappa_h \Delta_h(s_h, \hat{\pi}_h^k(s_h)) \middle| \mathcal{F}_{k-1}, s_1^k \right] \\
&\leq 2 \sum_{k=1}^K \sum_{h=1}^H \kappa_h \Delta_h(s_h^k, \hat{\pi}_h^k(s_h^k)) + \frac{16}{3} \tau^{-1} \log \log_2(K\kappa/\delta) + 2,
\end{aligned}$$

where the first equation is by Lemma B.1, the first inequality is by Assumption 4.2, and the second inequality is by Lemma A.5. □

The following lemma bounds the number of times a sub-optimality gap $\Delta_h(s_h^k, \hat{\pi}_h^k(s_h^k)$ exceeds a certain threshold.

Lemma B.4. *Under Assumption 4.1-4.2, for any $\Delta > 0$, w.p.a.l. $1 - 3\delta$, for any $h \in [H]$, we have:*

$$\sum_{k=1}^{K} \mathbb{1}\{\Delta_h(s_h^k, \hat{\pi}_h^k(s_h^k)) \geq \Delta\} \lesssim \frac{d^3 H^2 \iota^{-2}}{\Delta^2} \log^3(dKH/\delta).$$

Proof of Lemma B.4. Define $K' = \sum_{k=1}^{K} \mathbb{1}\{\Delta_h(s_h^k, \hat{\pi}_h^k(s_h^k)) \geq \Delta\}$. Note that K' is the number of episodes k where $\Delta_h(s_h^k, a_h^k)$ is bounded below by Δ. Define $\{k_i\}_{i \in [K']}$ such episodes, i.e. $k_i = \min\{k \in [K] : k \geq k_{i-1}, \Delta_h(s_h^k, a_h^k) \geq \Delta\}$. Then, we have:

$$\sum_{i=1}^{K'} \Delta_h(s_h^{k_i}, a_h^{k_i}) \geq K'\Delta.$$

Thus, with probability at least $1 - 3\delta$, for any $h \in [H]$, we have

$$\begin{aligned}
&\sum_{i=1}^{K'} \Delta_h(s_h^{k_i}, \hat{\pi}_h^{k_i}(s_h^{k_i})) = \sum_{i=1}^{K'} V_h^*(s_h^{k_i}) - Q_h^*(s_h^{k_i}, \hat{\pi}_h^{k_i}(s_h^{k_i})) \leq \sum_{i=1}^{K'} V_h^*(s_h^{k_i}) - Q_h^{\hat{\pi}^{k_i}}(s_h^{k_i}, \hat{\pi}_h^{k_i}(s_h^{k_i})) \\
&= \sum_{i=1}^{K'} V_h^*(s_h^{k_i}) - V_h^{\hat{\pi}^{k_i}}(s_h^{k_i}) \\
&\leq 2 \sum_{h'=h}^{H} \sum_{i=1}^{K'} \mathbb{E}_{\pi^*} \left[b_{h'}^{k_i}(s_{h'}, a_{h'}) | s_h^{k_i} \right] \\
&= 2 \sum_{h'=h}^{H} \sum_{i=1}^{K'} \beta_{k_i}(\delta) \mathbb{E}_{\pi^*} \left[\|\phi_{h'}(s_{h'}, a_{h'})\|_{(\Sigma_{h'}^{k_i})^{-1}} | s_h^{k_i} \right] \\
&\leq 2\beta_{K'}(\delta) \sum_{h'=h}^{H} \sum_{i=1}^{K'} \frac{d_{h'}^*(s_{h'}^{k_i}, a_{h'}^{k_i})}{d_{h'}^{\mu}(s_{h'}^{k_i}, a_{h'}^{k_i})} \|\phi_{h'}(s_{h'}^{k_i}, a_{h'}^{k_i})\|_{(\Sigma_{h'}^{k_i})^{-1}} + 2\beta_{K'}(\delta) \sum_{h'=h}^{H} \sqrt{\log(1/\delta)} \sqrt{\sum_{i=1}^{K'} \left(\frac{d_{h'}^*(s_{h'}^{k_i}, a_{h'}^{k_i})}{d_{h'}^{\mu}(s_{h'}^{k_i}, a_{h'}^{k_i})} \right)^2} \\
&\leq 2\beta_{K'}(\delta) \sum_{h'=h}^{H} \kappa_{h'} \sqrt{2K'd \log(1 + K'/d)} + 2\beta_{K'}(\delta) \sqrt{K' \log(1/\delta)} \sum_{h'=h}^{H} \kappa_{h'} \\
&\leq 2\kappa\beta_{K'}(\delta)(\sqrt{2K'd \log(1 + K'/d)} + 2\sqrt{K' \log(1/\delta)}) \\
&\lesssim \kappa H d^{3/2} K'^{1/2} \log^{3/2}(dK'H/\delta)
\end{aligned}$$

where the second inequality is by Lemma A.2, the third equality is by Lemma F.1, the third inequality is by Lemma A.7, the fourth inequality is by Lemma A.8. Thus, we have:

$$K' \lesssim \frac{d^3 H^2 \kappa^2}{\Delta^2} \log^3(dKH/\delta).$$

□

Next we bound the total sub-optimality gaps accumulated over K episodes under μ.

Lemma B.5. *Under Assumption 4.1-4.2-4.3, with probability at least $1 - 3\log_2(H/\Delta_{\min})\delta$, for any $h \in [H]$,*

$$\sum_{k=1}^{K} \Delta_h(s_h^k, \hat{\pi}_h^k(s_h^k)) \lesssim \frac{d^3 H^2 \kappa^2}{\Delta_{\min}} \log^3(dKH/\delta).$$

Proof of Lemma B.5. Let $m = \log_2(H/\Delta_{\min})$. As $\Delta_h(s_h^k, \hat{\pi}_h^k(s_h^k)) \in [0, H]$, and $\Delta_h(s_h^k, \hat{\pi}_h^k(s_h^k)) = 0$ if $\Delta_h(s_h^k, \hat{\pi}_h^k(s_h^k) < \Delta_{\min}$, we have:

$$
\begin{aligned}
\sum_{k=1}^{K} \Delta_h(s_h^k, \hat{\pi}_h^k(s_h^k)) &\leq \sum_{k=1}^{K} \sum_{i=1}^{m} \mathbb{1}\{2^{i-1}\Delta_{\min} \leq \Delta_h(s_h^k, \hat{\pi}_h^k(s_h^k)) < 2^i \Delta_{\min}\} 2^i \Delta_{\min} \\
&\leq \sum_{i=1}^{m} 2^i \Delta_{\min} \sum_{k=1}^{K} \mathbb{1}\{\Delta_h(s_h^k, \hat{\pi}_h^k(s_h^k)) \geq 2^{i-1}\Delta_{\min}\} \\
&\lesssim \sum_{i=1}^{m} 2^i \Delta_{\min} \frac{d^3 H^2 \kappa^2}{(2^{i-1}\Delta_{\min})^2} \log^3(dKH/\delta) \\
&\lesssim \frac{d^3 H^2 \kappa^2}{\Delta_{\min}} \log^3(dKH/\delta),
\end{aligned}
$$

where the first inequality is the peeling argument and the third inequality is by Lemma B.4. $\square$

Theorem 2 is a direct combination of Lemma B.3 and Lemma B.5 via union bound.

C Proof of Theorem 4

Let Ξ_h be the set of all trajectories $\tau_h := (s_1, a_1, \dots, s_h, a_h)$ induced by the underlying MDP and some policy π, i.e. $s_1 \sim d_1, a_i \sim \hat{\pi}_i(\cdot|s_i), s_{i+1} \sim \mathbb{P}_i(\cdot|s_i, a_i)$. Let

$$
\mathcal{E}_h^k = \{\tau_h = (s_1, a_1, \dots, s_h, a_h) \in \Xi_h : \forall i \in [h], a_i = \pi_i^*(s_i) = \hat{\pi}_h(s_i)\},
$$

be the set of all h-length trajectories $(s_1, a_1, \dots, s_h, a_h)$ at which $\hat{\pi}^k$ and π^* agree on up to step h.

C.1 Support lemmas

Next we show that the probability that π^* and $\hat{\pi}^k$ do not agree on a h-length trajectory is controlled by the sub-optimality and the minimum value gap.

Lemma C.1. *Under Assumption 4.3-4.4.1, for any $(k,h) \in [K] \times [H]$, if $f(k) \geq \sum_{t=1}^{k} \mathrm{SubOpt}(\hat{\pi}^t)$, we have:*

$$
\sum_{t=1}^{k} \mathbb{E}_{\tau_h \sim d^{\hat{\pi}^t}} \left[\mathbb{1}\{\tau_h \notin \mathcal{E}_h^t\} | \mathcal{F}_{t-1} \right] \leq \frac{1}{\Delta_{\min}} f(k).
$$

Proof of Lemma C.1. We have:

$$
\begin{aligned}
\sum_{t=1}^{k} \mathbb{E}_{\tau_h \sim d^{\hat{\pi}^t}} \left[\mathbb{1}\{\tau_h \notin \mathcal{E}_h^t\} | \mathcal{F}_{t-1} \right] &\leq \sum_{t=1}^{k} \sum_{i=1}^{h} \mathbb{E}_{(s_i, a_i) \sim d_i^{\hat{\pi}^t}} \left[\mathbb{1}\{a_i \neq \pi_i^*(s_i)\} | \mathcal{F}_{t-1} \right] \\
&= \sum_{t=1}^{k} \sum_{i=1}^{h} \mathbb{E}_{(s_i, a_i) \sim d_i^{\hat{\pi}^t}} \left[\mathbb{1}\{\Delta_i(s_i, a_i) \geq \Delta_{\min}\} | \mathcal{F}_{t-1} \right] \\
&\leq \sum_{t=1}^{k} \sum_{i=1}^{h} \mathbb{E}_{(s_i, a_i) \sim d_i^{\hat{\pi}^t}} \left[\frac{\Delta_i(s_i, a_i)}{\Delta_{\min}} \middle| \mathcal{F}_{t-1} \right] \\
&= \frac{1}{\Delta_{\min}} \sum_{t=1}^{k} \mathbb{E}_{\hat{\pi}^t} \left[\sum_{i=1}^{h} \Delta_i(s_i, a_i) \middle| \mathcal{F}_{t-1} \right]
\end{aligned}
$$

$$\leq \frac{1}{\Delta_{\min}} \sum_{t=1}^{k} \mathbb{E}_{\hat{\pi}^t} \left[\sum_{i=1}^{H} \Delta_i(s_i, a_i) \middle| \mathcal{F}_{t-1} \right]$$
$$= \frac{1}{\Delta_{\min}} \sum_{t=1}^{k} \mathrm{SubOpt}(\hat{\pi}^t)$$
$$\leq \frac{1}{\Delta_{\min}} f(k)$$

where the first equation is by Assumption 4.3 and Assumption 4.4.1, and the last equation is by Lemma B.1. ☐

The following lemma lower-bounds the empirical accumulated covariance matrix Σ_h^{k+1} by the covariance matrix at the optimal actions.

Lemma C.2. *Under Assumption 4.1-4.3-4.4.1, for any $(k, h) \in [K] \times [H]$, with probability at least $1 - \delta$, we have:*

$$\Sigma_h^{k+1} \succeq \lambda I + \kappa_{1:h}^{-1} k \Sigma_h^* - \kappa_{1:h}^{-1} I \frac{1}{\Delta_{\min}} \sum_{t=1}^{k} \mathrm{SubOpt}(\hat{\pi}^t) - \frac{2}{3} \log(d/\delta) I - \sqrt{2k \log(d/\delta)} I.$$

Proof. Let $Z_t := \mathbb{1}\{\tau_h^t \in \mathcal{E}_h^t\} \phi_h^*(s_h^t) \phi_h^*(s_h^t)^T - \mathbb{E}\left[\mathbb{1}\{\tau_h^t \in \mathcal{E}_h^t\} \phi_h^*(s_h^t) \phi_h^*(s_h^t)^T | \mathcal{F}_{t-1}\right]$ and $\kappa_{1:h} := \prod_{i=1}^{h} \kappa_i$. We have:

$$\Sigma_h^{k+1} - \lambda I = \sum_{t=1}^{k} \phi_h(s_h^t, a_h^t) \phi_h(s_h^t, a_h^t)^T$$
$$\succeq \sum_{t=1}^{k} \mathbb{1}\{\tau_h^t \in \mathcal{E}_h^t\} \phi_h(s_h^t, a_h^t) \phi_h(s_h^t, a_h^t)^T$$
$$\overset{(a)}{=} \sum_{t=1}^{k} \mathbb{1}\{\tau_h^t \in \mathcal{E}_h^t\} \phi_h^*(s_h^t) \phi_h^*(s_h^t)^T$$
$$= \sum_{t=1}^{k} \mathbb{E}_{\tau_h \sim d^\mu} \left[\mathbb{1}\{\tau_h \in \mathcal{E}_h^t\} \phi_h^*(s_h) \phi_h^*(s_h)^T | \mathcal{F}_{t-1}\right] + \sum_{t=1}^{k} Z_t$$
$$\overset{(b)}{\succeq} \sum_{t=1}^{k} \kappa_{1:h}^{-1} \mathbb{E}_{\tau_h \sim d^*} \left[\mathbb{1}\{\tau_h \in \mathcal{E}_h^t\} \phi_h^*(s_h) \phi_h^*(s_h)^T | \mathcal{F}_{t-1}\right] + \sum_{t=1}^{k} Z_t$$
$$= \sum_{t=1}^{k} \kappa_{1:h}^{-1} \mathbb{E}_{\tau_h \sim d^*} \left[\phi_h^*(s_h) \phi_h^*(s_h)^T | \mathcal{F}_{t-1}\right] - \sum_{t=1}^{k} \kappa_{1:h}^{-1} \mathbb{E}_{\tau_h \sim d^*} \left[\mathbb{1}\{\tau_h \notin \mathcal{E}_h\} \phi_h^*(s_h) \phi_h^*(s_h)^T | \mathcal{F}_{t-1}\right] + \sum_{t=1}^{k} Z_t$$
$$\overset{(c)}{\succeq} \sum_{t=1}^{k} \kappa_{1:h}^{-1} \Sigma_h^* - \kappa_{1:h}^{-1} I \sum_{t=1}^{k} \mathbb{E}_{\tau_h \sim d^*} \left[\mathbb{1}\{\tau_h \notin \mathcal{E}_h^t\} | \mathcal{F}_{t-1}\right] + \sum_{t=1}^{k} Z_t$$
$$= \kappa_{1:h}^{-1} k \Sigma_h^* - \kappa_{1:h}^{-1} I \sum_{t=1}^{k} \mathbb{E}_{\tau_h \sim d^*} \left[1 - \mathbb{1}\{\tau_h \in \mathcal{E}_h^t\} | \mathcal{F}_{t-1}\right] + \sum_{t=1}^{k} Z_t$$
$$= \kappa_{1:h}^{-1} k \Sigma_h^* - \kappa_{1:h}^{-1} I \sum_{t=1}^{k} (1 - \mathbb{E}_{\tau_h \sim d^*} \left[\mathbb{1}\{\tau_h \in \mathcal{E}_h^t\} | \mathcal{F}_{t-1}\right]) + \sum_{t=1}^{k} Z_t$$
$$\overset{(d)}{=} \kappa_{1:h}^{-1} k \Sigma_h^* - \kappa_{1:h}^{-1} I \sum_{t=1}^{k} (1 - \mathbb{E}_{\tau_h \sim d^{\hat{\pi}^t}} \left[\mathbb{1}\{\tau_h \in \mathcal{E}_h^t\} | \mathcal{F}_{t-1}\right]) + \sum_{t=1}^{k} Z_t$$
$$= \kappa_{1:h}^{-1} k \Sigma_h^* - \kappa_{1:h}^{-1} I \underbrace{\sum_{t=1}^{k} \mathbb{E}_{\tau_h^t \sim d^{\hat{\pi}^t}} \left[\mathbb{1}\{\tau_h \notin \mathcal{E}_h^t\} | \mathcal{F}_{t-1}\right]}_{(i)} + \underbrace{\sum_{t=1}^{k} Z_t}_{(ii)}$$

where (a) is by the definition of $\mathcal{E}_h^k$, (b) is by that under Assumption 4.1, we have

$$\frac{d^\mu(\tau_h)}{d^*(\tau_h)} = \frac{d_1^\mu(s_1,a_1)\mathbb{P}_1(s_2|s_1,a_1)\dots\mathbb{P}_{h-1}(s_h|s_{h-1},a_{h-1})d_h^\mu(s_h,a_h)}{d_1^*(s_1,a_1)\mathbb{P}_1(s_2|s_1,a_1)\dots\mathbb{P}_{h-1}(s_h|s_{h-1},a_{h-1})d_h^*(s_h,a_h)} = \prod_{i=1}^{h}\frac{d_i^\mu(s_i,a_i)}{d_i^*(s_i,a_i)} \ge \kappa_{1:h}^{-1},$$

(c) is by that $\phi_h^*(s_h)\phi_h^*(s_h)^T \preceq I \cdot \|\phi_h^*(s_h)\phi_h^*(s_h)^T\| \le I\|\phi_h^*(s_h)\|_2^2 = I$, and that

$$\mathbb{E}_{\tau_h^t\sim d^*}\left[\phi_h^*(s_h^t)\phi_h^*(s_h^t)^T|\mathcal{F}_{t-1}\right] = \mathbb{E}_{(s_h,a_h)\sim d_h^*}\left[\phi_h^*(s_h)\phi_h^*(s_h)^T\right] = \Sigma_h^*,$$

and (d) is by that $d^*(\tau_h|\mathcal{E}_h^t) = d^{\hat{\pi}^t}(\tau_h|\mathcal{E}_h^t)$.

Term (i) is bounded by Lemma C.1. For term (ii), note that Z_t is $\mathcal{F}_t$-measurable, $\mathbb{E}\left[Z_t|\mathcal{F}_{t-1}\right] = 0$, and $\|Z_t\|_2 \le 1$. Thus, by Lemma E.14, with probability at least $1-\delta$, we have:

$$(ii) = \sum_{t=1}^{k} Z_t \ge -\frac{2}{3}\log(d/\delta)I - \sqrt{2k\log(d/\delta)}I.$$

□

Lemma C.3. *Let λ_h^+ be the smallest positive eigenvalue of Σ_h^*, $\kappa_{1:h} := \prod_{i=1}^h \kappa_h$, and define*

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

Under Assumption 4.1-4.3-4.4.1, w.p.a.l. $1-2\delta$, for any $h \in [H]$, any $k \ge \max_{h\in[H]}\bar{k}_h$, and any $v \in col(\Sigma_h^*)$ such that $\|v\|_2 \le 1$, we have:

$$\|v\|_{(\Sigma_h^k)^{-1}} = \mathcal{O}\left(\sqrt{\frac{\kappa_{1:h}}{(\lambda_h^+)^3k}}\right).$$

Proof. By Lemma C.2 and the union bound, with probability at least $1-\delta$, for any $(k,h) \in [K]\times[H]$, we have: $\Sigma_h^{k+1} \succeq A_h^k$ where

$$A_h^k := \lambda I + \kappa_{1:h}^{-1}k\Sigma_h^* - \kappa_{1:h}^{-1}I\frac{1}{\Delta_{\min}}f(k) - \frac{2}{3}\log(dKH/\delta)I - \sqrt{2k\log(dKH/\delta)}I,$$

where $f(k)$ is any upper bound of $\sum_{t=1}^k \mathrm{SubOpt}(\hat{\pi}^t)$. We can choose $f(k) = \tilde{\mathcal{O}}\left(\kappa H d^{3/2}\sqrt{k}\right)$ by Corollary 1, or $f(k) = \tilde{\mathcal{O}}\left(\frac{d^3H^2\kappa^3}{\Delta_{\min}}\right)$ by Theorem 2. We fix h. Let $0 \le \lambda_1 \le \ldots \le \lambda_d$ be the eigenvalues of Σ_h^* with the corresponding orthonormal vectors $u_1,\ldots,u_d$, and let λ_h^+ the smallest positive eigenvalue of Σ_h^*. Then, A_h^k has the eigenvalues $\lambda_1' \le \ldots \le \lambda_d'$ with the same corresponding orthonormal vectors $u_1,\ldots,u_d$, where:

$$\lambda_i' := 1 + \kappa_{1:h}^{-1}k\lambda_i - \kappa_{1:h}^{-1}\frac{1}{\Delta_{\min}}f(k) - \frac{2}{3}\log(dkH/\delta) - \sqrt{2k\log(dkH/\delta)}$$

It is easy to verify that if $k \ge k_h$, we have:

$$1 - \kappa_{1:h}^{-1}\frac{1}{\Delta_{\min}}f(k) - \frac{2}{3}\log(dkH/\delta) < \sqrt{2k\log(dkH/\delta)}, \text{ and}$$
$$1 + \kappa_{1:h}^{-1}k\lambda_h^+ - \kappa_{1:h}^{-1}\frac{1}{\Delta_{\min}}f(k) - \frac{2}{3}\log(dkH/\delta) - \sqrt{2k\log(dkH/\delta)} > 0.$$

Thus, for any $k \geq k_h$, $\lambda'_i \neq 0, \forall i \in [d]$, i.e. A_h^k is invertible, we have:

$$\bar{\lambda}_h^+ := 1 + \kappa_{1:h}^{-1} k \lambda_h^+ - \kappa_{1:h}^{-1} \frac{1}{\Delta_{\min}} f(k) - \frac{2}{3}\log(dkH/\delta) - \sqrt{2k\log(dkH/\delta)}$$

is the smallest positive eigenvalue of A_h^k, and hence for any $v \in \mathrm{span}\{\phi_h^*(s) | d_h^*(s) > 0\}$, such that $\|v\|_2 \leq 1$, let $x = v/\|v\|_2$. We have:

$$\|v\|_{(\Sigma_h^{k+1})^{-1}} \leq \|x\|_{(\Sigma_h^{k+1})^{-1}} \leq \|x\|_{(A_h^k)^{-1}} \leq \frac{\lambda'_d}{\bar{\lambda}_h^+} \frac{1}{\|x\|_{A_h^k}} \leq \frac{\lambda'_d}{(\bar{\lambda}_h^+)^{3/2}} \frac{1}{\|x\|_2} = \frac{\lambda'_d}{(\bar{\lambda}_h^+)^{3/2}} = \mathcal{O}\left(\sqrt{\frac{\kappa_{1:h}}{(\lambda_h^+)^3 k}}\right)$$

where the first inequality is by $\|v\|_2 \leq 1$, the second inequality is by $(\Sigma_h^{k+1})^{-1} \preceq (A_h^k)^{-1}$, the third inequality is by Lemma E.8, and the fourth inequality is by Lemma E.7, $\|v\|_{A_h^k} \geq \|v\|_2\sqrt{\bar{\lambda}_h^+}$. Choosing $\bar{k} = \max_h \bar{k}_h$ completes the proof. □

C.2 Proof of Theorem 4

Proof of Theorem 4. Let $\mathcal{E}$ be the event that the inequalities in Corollary 1, Theorem 2, and Lemma C.3 hold simultaneously. We now consider event $\mathcal{E}$ for the rest of the proof. Consider any state $s_h \in \mathcal{S}_h^\mu$. By Assumption 4.4.2, we have $\phi_h^*(s_h) \in \mathrm{col}(\Sigma_h^*)$. Thus, by Lemma C.3, if $k \geq \bar{k}_h$

$$2\beta_k(\delta)\|\phi_h(s_h, a_h)\|_{(\Sigma_h^k)^{-1}} = \tilde{\mathcal{O}}\left(dH\sqrt{\frac{\kappa_{1:h}}{(\lambda_h^+)^3 k}}\right).$$

We choose k such that $k \geq \max_h \bar{k}_h$ and

$$\tilde{\Omega}\left(dH\sqrt{\frac{\kappa_{1:h}}{(\lambda_h^+)^3 k}}\right) \leq \frac{\Delta_{\min}}{H},$$

i.e.

$$k \geq \tilde{\Omega}\left(\frac{d^2 H^4 \kappa_{1:h}}{\Delta_{\min}^2 (\lambda_h^+)^3}\right), \forall h.$$

Then, we have:

$$\begin{aligned}
\Delta_h(s_h, \hat{\pi}^k(s_h)) &= V_h^*(s_h) - Q^*(s_h, \hat{\pi}^k(s_h)) \\
&\leq V_h^*(s_h) - Q^{\hat{\pi}^k}(s_h, \hat{\pi}^k(s_h)) \\
&\leq 2\beta_k(\delta)\mathbb{E}_{\pi^*}\left[\sum_{h'=h}^H \|\phi_h(s_h, a_h)\|_{(\Sigma_{h'}^k)^{-1}} | \mathcal{F}_{k-1}, s_h\right] \\
&< (H - h + 1)\frac{\Delta_{\min}}{H} \leq \Delta_{\min}.
\end{aligned}$$

Thus, $\Delta_h(s_h, \hat{\pi}^k(s_h)) = 0, \forall h$. Therefore, for any initial state $s_1 \sim d_1$, we have:

$$\begin{aligned}
\mathrm{SubOpt}(\hat{\pi}^k; s_1) &= \mathbb{E}_{\hat{\pi}^k}\left[\sum_{h=1}^H \Delta_h(s_h, a_h)\middle| \mathcal{F}_{k-1}, s_1\right] \\
&= \mathbb{E}_{\hat{\pi}^k}\left[\sum_{h=1}^H \Delta_h(s_h, \hat{\pi}_h^k(s_h))\middle| \mathcal{F}_{k-1}, s_1\right] \\
&\leq \mathbb{E}_{\mu}\left[\sum_{h=1}^H \kappa_h \Delta_h(s_h, \hat{\pi}_h^k(s_h))\middle| \mathcal{F}_{k-1}, s_1\right] = 0
\end{aligned}$$

where the first equation is by Lemma B.1 and the inequality is by Lemma B.2 and Assumption 4.2.

□

D Proofs of the information-theoretic lower bound

In this section, we give the proof of Theorem 3. To prove the lower bound, we construct the hard MDP instances introduced by (Jin et al., 2021). The key difference is that in our construction, we need to carefully design the behavior policy μ to incorporate the optimal-policy concentrability $\{\kappa_h\}_{h\in[H]}$ and the minimum positive action gap $\Delta_{\min}$ into the lower bound.

D.1 Construction of a hard instance

We construct a class of MDPs parameterized by $M(p_1, p_2)$ with horizon $H \geq 2$, action space $\mathcal{A} = \{b_i\}_{i=1}^A$ (where $A \geq 2$), state space $\mathcal{S} = \{x_0, x_1, x_2\}$, initial state distribution $d_1(x_0) = 1$, transition kernels $\mathbb{P}_1(x_1|x_0, p_i) = p_i$, $\mathbb{P}_1(x_2|x_0, p_i) = 1 - p_i, \forall i \in [A]$ where $p_i := \min\{p_1, p_2\}, \forall i \geq 3$, $\mathbb{P}_h(x_1|x_1, a) = 1, \forall (h, a) \in [H] \times \mathcal{A}$, and reward functions $r_h(s_h, a_h) = 1\{s_h = x_1, h \geq 2\}$. It is not hard to see that the optimal action at the first stage is b_{i^*} where $i^* = \arg\max\{p_i : i \in \{1, 2\}\}$ and the optimal action at any stage $h \geq 2$ is any action $a \in \mathcal{A}$. The diagram of $M(p_1, p_2)$ is depicted in Figure 2. It is easy to see that this MDP satisfies both the definition of a linear MDP (see Definition 1) and of a mixture linear MDP (see Definition 3). By direct computation, we have:

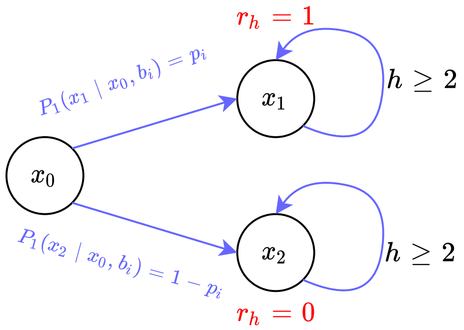

Figure 2: The diagram of the hard MDP introduced by (Jin et al., 2021).

$$\begin{cases} V_1^*(x_0) & = \max\{p_1, p_2\}(H-1), \\ Q_1^*(x_0, b_i) & = p_i(H-1), \forall i \in [A], \\ Q_h^*(x_1, a) & = H - h, \forall (h, a) \in \{2, \dots, H\} \times \mathcal{A}, \\ Q_h^*(x_2, a) & = 0, \forall (h, a) \in \{2, \dots, H\} \times \mathcal{A}. \end{cases}$$

$$\begin{cases} V_1^\pi(x_0) & = \sum_{i=1}^A \pi_1(b_i|x_0)p_i(H-1), \\ Q_1^\pi(x_0, b_i) & = p_i(H-1), \forall i \in [A], \\ Q_h^\pi(x_1, a) & = H - h, \forall (h, a) \in \{2, \dots, H\} \times \mathcal{A}, \\ Q_h^\pi(x_2, a) & = 0, \forall (h, a) \in \{2, \dots, H\} \times \mathcal{A}. \end{cases}$$

$$\begin{cases} d_1^*(x_0) & = 1, \\ d_h^*(x_1) & = \max\{p_1, p_2\}, \forall h \geq 2, \\ d_h^*(x_2) & = 1 - \max\{p_1, p_2\}, \forall h \geq 2. \end{cases}$$

$$\begin{cases} d_1^\pi(x_0) & = 1, \\ d_h^\pi(x_1) & = \sum_{i=1}^A \pi_1(b_i|x_0)p_i, \\ d_h^\pi(x_2) & = \sum_{i=1}^A \pi_1(b_i|x_0)(1-p_i). \end{cases}$$

We also have $\Delta_{\min} = |p_1 - p_2|(H-1)$.

D.2 Proof of Theorem 3

Proof of Theorem 3. We consider two MDP instances $M_1 := M(p^*, p)$ and $M_2 := M(p, p^*)$ (the parameterization is defined in Subsection D.1) where $p^* > p$, whose optimal actions in the first stage are b_1 and b_2, respectively. The intuition for the hardness of these two instances is that any policy is sub-optimal in at least one of the instances. We have:

$$\text{SubOpt}(\pi; M_1) = (H-1)(p^* - p)(1 - \pi_1(b_1|x_0)),$$
$$\text{SubOpt}(\pi; M_2) = (H-1)(p^* - p)(1 - \pi_1(b_2|x_0)).$$

Consider any policy $\pi = \text{Algo}(\mathcal{D})$ and let $a_1 \sim \pi_1(\cdot|x_0)$ (note that a_1 is a random variable). We have:

$$\begin{aligned}
&2 \max_{l \in \{1,2\}} \mathbb{E}_{\mathcal{D}\sim M_l} [\text{SubOpt}(\text{Algo}(\mathcal{D}); M_l)] \\
&\geq \mathbb{E}_{\mathcal{D}\sim M_l} [\text{SubOpt}(\text{Algo}(\mathcal{D}); M_1)] + \mathbb{E}_{\mathcal{D}\sim M_2} [\text{SubOpt}(\text{Algo}(\mathcal{D}); M_2)] \\
&= (H-1)(p^* - p) \left(\mathbb{E}_{\mathcal{D}\sim M_1} [1 - \pi_1(b_1|x_0)] + \mathbb{E}_{\mathcal{D}\sim M_2} [1 - \pi_1(b_2|x_0)]\right) \\
&\geq (H-1)(p^* - p) \left(\mathbb{E}_{\mathcal{D}\sim M_1} [1 - \pi_1(b_1|x_0)] + \mathbb{E}_{\mathcal{D}\sim M_2} [\pi_1(b_1|x_0)]\right) \\
&= (H-1)(p^* - p) \left(\mathbb{E}_{\mathcal{D}\sim M_1} [1\{a_1 \neq b_1\}] + \mathbb{E}_{D\sim M_2} [1\{a_1 = b_1\}]\right) \\
&\geq (H-1)(p^* - p)(1 - \text{TV}(P_{M_1}, P_{M_2})) \\
&\geq (H-1)(p^* - p)(1 - \sqrt{KL(P_{\mathcal{D}\sim M_1} \| P_{\mathcal{D}\sim M_2})/2}),
\end{aligned}$$

where the third inequality is by the definition of the total variation distance $\text{TV}(P, Q) = \sup\{|P(B) - Q(B)| : \forall B \text{ is measurable}\}$, and the last inequality is by Donsker's inequality.

Construction of behavior policy. To construct the behaviour policy μ that satisfies $\sup_{h,s_h,a_h} \frac{d^*_h(s_h,a_h)}{d^\mu_h(s_h,a_h)} \leq \kappa_h, \forall h \in [H]$ in both $M(p^*, p)$ and $M(p, p^*)$, we consider $\mu_h(a|x_i) = \frac{1}{A}, \forall (h, a, i) \in \{2, \ldots, H\} \times \mathcal{A} \times \{1, 2\}$. We also set $\mu_1(b_1|x_0) = \mu_1(b_2|x_0) = q$. By direct computation, we have:

$$\begin{cases}
\max_{s_1,a_1} \frac{d_1^{M_i,*}(s_1,a_1)}{d_1^{M_i,\mu}(s_1,a_1)} & = \frac{1}{q} \\
\max_{s_h,a_h} \frac{d_h^{M_i,*}(s_h,a_h)}{d_h^{M_i,\mu}(s_h,a_h)} & \leq \max\left\{\frac{p^*}{q(p^*+p)}, \frac{1-p^*}{q(2-p^*-p)}\right\} = \frac{p^*}{q(p^*+p)} \leq \frac{1}{q}
\end{cases}$$

As $\kappa_h \geq 2$, we set $q = \frac{1}{\min_h \kappa_h} \in (0, \frac{1}{2}]$, and thus we have $\sup_{h,s_h,a_h} \frac{d^*_h(s_h,a_h)}{d^\mu_h(s_h,a_h)} \leq \kappa_h, \forall h \in [H]$.

Computing $KL(\mathbb{P}_{\mathcal{D}\sim M_1} \| \mathbb{P}_{\mathcal{D}\sim M_2})$. We consider dataset $\mathcal{D} = \{(s_h^t, a_h^t, r_h^t)\}_{h\in[H]}^{t\in[K]}$ such that $s_1^t = x_0, \forall t \in [K]$ and $\mathcal{D}$ agrees with the class $\{M(p_1, p_2) : p_1 \neq p_2\}$ and behavior policy μ. Under $M(p_1, p_2)$, we have $r_1^t = 0, \forall t \in [K]$ and $s_h^t = s_2^t, r_h^t = r_2^t, \forall t \in [K], \forall h \geq 2$. Thus, $\mathcal{D}' := \{(s_1^t, a_1^t, s_2^t, r_2^t)\}^{t\in[K]}$ is as informative as $\mathcal{D}$. [5] Thus, $KL(\mathbb{P}_{\mathcal{D}\sim M_1} \| \mathbb{P}_{\mathcal{D}\sim M_2}) = KL(\mathbb{P}_{\mathcal{D}'\sim M_1} \| \mathbb{P}_{\mathcal{D}'\sim M_2})$.

Let $n_l := \sum_{t=1}^K 1\{a_1^t = b_l\}$ be the number of episodes in $\mathcal{D}'$ with the first-stage action b_l and $Z_l = \{r_2^t : a_2^t = b_l, t \in [K]\}$ be the set of the second-stage reward in $\mathcal{D}'$ in such episodes. We have:

$$\mathbb{P}_{M_1}(\mathcal{D}') = \mu_1^{n_1}(b_1|x_0)(p^*)^{\sum_{z\in Z_1} z}(1-p^*)^{n_1 - \sum_{z\in Z_1} z} \prod_{i\neq 1} \mu_1^{n_i}(b_i|x_0)p^{\sum_{z\in Z_i} z}(1-p)^{n_i - \sum_{z\in Z_i} z},$$

[5] In the sense that knowing $\mathcal{D}'$ implies $\mathcal{D}$.

$$\mathbb{P}_{M_2}(\mathcal{D}') = \mu_1^{n_1}(b_2|x_0)(p^*)^{\sum_{z\in Z_2} z}(1-p^*)^{n_2-\sum_{z\in Z_2} z} \prod_{i\neq 2} \mu_1^{n_i}(b_i|x_0) p^{\sum_{z\in Z_i} z}(1-p)^{n_i - \sum_{z\in Z_i} z}.$$

Thus, we have:

$$\log \frac{\mathbb{P}_{M_1}(\mathcal{D}')}{\mathbb{P}_{M_2}(\mathcal{D}')} = \delta_{1,2} \log \frac{p^*(1-p)}{p(1-p^*)} + (n_1 - n_2) \log \frac{1-p^*}{1-p},$$

where $\delta_{1,2} := \sum_{z\in Z_1} z - \sum_{z\in Z_2} z$. Thus, we have:

$$\begin{aligned}
KL(\mathbb{P}_{\mathcal{D}'\sim M_1} \| \mathbb{P}_{\mathcal{D}'\sim M_2}) &= \mathbb{E}_{\mathcal{D}'\sim M_1}\left[\log \frac{\mathbb{P}_{M_1}(\mathcal{D}')}{\mathbb{P}_{M_2}(\mathcal{D}')}\right] \\
&= \mathbb{E}_{\mathcal{D}'\sim M_1}[\delta_{1,2}] \log \frac{p^*(1-p)}{p(1-p^*)} + \mathbb{E}_{\mathcal{D}'\sim M_1}[n_1 - n_2] \log \frac{1-p^*}{1-p} \\
&= Kq(p^* - p) \log \frac{p^*(1-p)}{p(1-p^*)} \\
&= Kq(p^* - p) \log \left(1 + \frac{p^* - p}{p(1-p^*)}\right).
\end{aligned}$$

Construction of (p^*, p). Now we choose $p, p^* \in [\frac{1}{4}, \frac{3}{4}]$ and $p^* - p \leq \frac{1}{16}$. For such p^*, p, we have $\log\left(1 + \frac{p^*-p}{p(1-p^*)}\right) \leq \log(1 + 16(p^* - p)) \leq 16(p^* - p)$, where the last inequality is by that $\log(1+x) \leq x, \forall x \in (0,1]$. Thus, we have $KL(\mathbb{P}_{\mathcal{D}'\sim M_1} \| \mathbb{P}_{\mathcal{D}'\sim M_2}) \leq 16qK(p^* - p)^2$. Hence, we have:

$$\begin{aligned}
2 \max_{l\in\{1,2\}} \mathbb{E}_{\mathcal{D}\sim M_l}[\text{SubOpt}(\text{Algo}(\mathcal{D}); M_l)] &\geq (H-1)(p^* - p)(1 - \sqrt{KL(\mathbb{P}_{\mathcal{D}\sim M_1} \| \mathbb{P}_{\mathcal{D}\sim M_2})/2}) \\
&\geq (H-1)(p^* - p)(1 - 2(p^* - p)\sqrt{2qK}).
\end{aligned}$$

Note that we choose $q = 1/\kappa_{\min}$. Now, we simply set:

$$p^* - p = \frac{1}{4\sqrt{2}} \sqrt{\frac{\kappa_{\min}}{K}}.$$

Then, we have $1 - 2(p^* - p)\sqrt{2qK} = \frac{1}{2}$, $\Delta_{\min} = (p^* - p)(H-1) = \frac{H-1}{4\sqrt{2}}\sqrt{\frac{\kappa_{\min}}{K}}$, and thus

$$\Delta_{\min} = \frac{1}{32} \frac{(H-1)^2 \kappa_{\min}}{K \Delta_{\min}}.$$

Therefore, we have:

$$\begin{aligned}
\max_{l\in\{1,2\}} \mathbb{E}_{\mathcal{D}\sim M_l}[\text{SubOpt}(\text{Algo}(\mathcal{D}); M_l)] &\geq \frac{1}{2}(H-1)(p^* - p)(1 - 2(p^* - p)\sqrt{2qK}) \\
&= 2^{-7} \frac{(H-1)^2 \kappa_{\min}}{K \Delta_{\min}} = \Omega\left(\frac{H^2 \kappa_{\min}}{K \Delta_{\min}}\right).
\end{aligned}$$

□

E Auxiliary lemmas

E.1 MDPs

Lemma E.1 (Extended Value Difference (Cai et al., 2020, Section B.1))**.** *Let* $\pi = \{\pi_h\}_{h=1}^H$ *and* $\pi' = \{\pi'_h\}_{h=1}^H$ *be two arbitrary policies and let* $\{Q_h\}_{h=1}^H$ *be arbitrary functions* $\mathcal{S} \times \mathcal{A} \to \mathbb{R}$. *Let* $V_h := \langle Q_h, \pi_h \rangle$. *Then* $\forall s \in \mathcal{S}$, $\forall h \in [H]$,

$$V_h(s) - V_h^{\pi'}(s) = \sum_{i=h}^H \mathbb{E}_{\pi'}\left[\langle Q_i(s_i, \cdot), \pi_i(\cdot|s_i) - \pi'_i(\cdot|s_i)\rangle | s_h = s\right]$$

$$+ \sum_{i=h}^{H} \mathbb{E}_{\pi'} \left[Q_i(s_i, a_i) - T_i V_{i+1}(s_i, a_i) | s_h = s \right],$$

where $T_i V := r_i + P_i V$ *and* $\mathbb{E}_{\pi'}$ *is the expectation over the randomness of* $(s_h, a_h, \ldots, s_H, a_H)$ *induced by* π'*.*

Proof of Lemma E.1. Fix $h \in [H]$. Denote $\xi_i := Q_i - T_i V_{i+1}$. For $\forall i \in [h, H-1]$, $\forall s_h \in \mathcal{S}$, we have

$$\begin{aligned}
\mathbb{E}_{\pi'} \left[V_i(s_i) - V_i^{\pi'}(s_i) | s_h \right] &= \mathbb{E}_{\pi'} \left[\langle Q_i(s_i, \cdot), \pi_i(\cdot|s_i) \rangle - \langle Q_i^{\pi'}(s_i, \cdot), \pi_i'(\cdot|s_i) \rangle | s_h \right] \\
&= \mathbb{E}_{\pi'} \left[\langle Q_i(s_i, \cdot), \pi_i(\cdot|s_i) - \pi_i'(\cdot|s_i) \rangle + \langle Q_i(s_i, \cdot) - Q_i^{\pi'}(s_i, \cdot), \pi_i'(\cdot|s_i) \rangle | s_h \right] \\
&= \mathbb{E}_{\pi'} \left[\langle Q_i(s_i, \cdot), \pi_i(\cdot|s_i) - \pi_i'(\cdot|s_i) \rangle | s_h \right] \\
&+ \mathbb{E}_{\pi'} \left[\langle \xi_i(s_i, \cdot) + T_i V_{i+1}(s_i, \cdot) - (r_i(s_i, \cdot) + P_i V_{i+1}^{\pi'}(s_i, \cdot)), \pi_i'(\cdot|s_i) \rangle | s_h \right] \\
&= \mathbb{E}_{\pi'} \left[\langle Q_i(s_i, \cdot), \pi_i(\cdot|s_i) - \pi_i'(\cdot|s_i) \rangle | s_h \right] + \mathbb{E}_{\pi'} \left[\xi_i(s_i, a_i) | s_h \right] \\
&+ \mathbb{E}_{\pi'} \left[P_i(V_{i+1} - V_{i+1}^{\pi'})(s_i, a_i) | s_h \right] \\
&= \mathbb{E}_{\pi'} \left[\langle Q_i(s_i, \cdot), \pi_i(\cdot|s_i) - \pi_i'(\cdot|s_i) \rangle | s_h \right] + \mathbb{E}_{\pi'} \left[\xi_i(s_i, a_i) | s_h \right] \\
&+ \mathbb{E}_{\pi'} \left[V_{i+1}(s_{i+1}) - V_{i+1}^{\pi'}(s_{i+1}) | s_h \right].
\end{aligned}$$

Taking $\sum_{i=h}^{H}$ both sides of the last equation above completes the proof. □

Lemma E.2. *Let* $\hat{\pi} = \{\hat{\pi}_h\}_{h=1}^{H}$ *and* $\hat{Q}_h(\cdot, \cdot)$ *be arbitrary policy and Q-function. Let* $\hat{V}_h(s) = \langle \hat{Q}_h(s, \cdot), \hat{\pi}_h(\cdot|s) \rangle$ *and* $\zeta_h(s, a) := (T_h \hat{V}_{h+1})(s, a) - \hat{Q}_h(s, a)$*. For any policy* π *and* $h \in [H]$*, we have*

$$\begin{aligned}
V_h^{\pi}(s) - V_h^{\hat{\pi}}(s) &= \sum_{i=h}^{H} \mathbb{E}_{\pi} \left[\zeta_i(s_i, a_i) | s_h = s \right] - \sum_{i=h}^{H} \mathbb{E}_{\hat{\pi}} \left[\zeta_i(s_i, a_i) | s_h = s \right] \\
&+ \sum_{i=h}^{H} \mathbb{E}_{\pi} \left[\langle \hat{Q}_h(s_h, \cdot), \pi_h(\cdot|s_h) - \hat{\pi}_h(\cdot|s_h) \rangle | s_h = s \right].
\end{aligned}$$

Proof. We apply Lemma E.1 with $\pi = \hat{\pi}$, $\pi' = \hat{\pi}$, $Q_h = \hat{Q}_h$ and apply Lemma E.1 again with $\pi = \hat{\pi}$, $\pi' = \pi$, $Q_h = \hat{Q}_h$ and take the difference between two results to complete the proof. □

Lemma E.3. *For any* $0 \leq V(\cdot) \leq H$*, there exists a* $w_h \in \mathbb{R}^d$ *such that* $T_h V = \langle \phi_h, w_h \rangle$ *and* $\|w_h\|_2 \leq 2H\sqrt{d}$*. In addition, for any policy* $\pi \in \Pi$*,* $\exists w_h^{\pi} \in \mathbb{R}^d$ *s.t.* $Q_h^{\pi}(s, a) = \phi_h(s, a)^T w_h^{\pi}$ *with* $\|w_h^{\pi}\|_2 \leq 2(H - h + 1)\sqrt{d}$*.*

Proof. By definition,

$$T_h V = r_h + P_h V = \langle \phi, \theta_h \rangle + \langle \phi, \int_{\mathcal{S}} V(s) d\nu_h(s) \rangle = \langle \phi, w_h \rangle,$$

where $w_h = \theta_h + \int_{\mathcal{S}} V(s) d\nu_h(s)$. By the assumption of linear MDP,

$$\|w_h\|_2 = \|\theta_h + \int_{\mathcal{S}} V(s) d\nu_h(s)\|_2 \leq \|\theta_h\|_2 + \|\int_{\mathcal{S}} V(s) d\nu_h(s)\|_2 \leq \sqrt{d} + H\sqrt{d} \leq 2H\sqrt{d}.$$

The second part is similar with $V_h^{\pi} \leq H - h + 1$. □

Lemma E.4 (Bound on weights in algorithm)**.** *For any $(k,h) \in [K] \times [H]$, the weight $\hat{w}_h^k$ in Algorithm 1 satisfies:*

$$\|\hat{w}_h^k\|_2 \leq (H-h+1)\sqrt{dk/\lambda}.$$

Proof. For any $v \in \mathbb{R}^d$, we have

$$\begin{aligned}
|v^T \hat{w}_h^k| &= \left| v^T (\Sigma_h^k)^{-1} \sum_{t=1}^{k} \phi_h(s_h^t, a_h^t)(r_h^t + \hat{V}_{h+1}^k(s_{h+1}^t)) \right| \leq (H-h+1) \sum_{t=1}^{k} \left| v^T (\Sigma_h^k)^{-1} \phi_h(s_h^t, a_h^t) \right| \\
&\leq (H-h+1)\|v\|_{(\Sigma_h^k)^{-1}} \sum_{t=1}^{k} \|\phi_h(s_h^t, a_h^t)\|_{(\Sigma_h^k)^{-1}} \\
&\leq (H-h+1)\|v\|_{(\Sigma_h^k)^{-1}} \sqrt{k \sum_{t=1}^{k} \|\phi_h(s_h^t, a_h^t)\|^2_{(\Sigma_h^k)^{-1}}} \leq (H-h+1)\|v\|_2 \sqrt{\|(\Sigma_h^k)^{-1}\|} \cdot \sqrt{kd} \\
&\leq (H-h+1)\sqrt{kd/\lambda} \cdot \|v\|_2,
\end{aligned}$$

where the penultimate inequality is due to that

$$\begin{aligned}
\sum_{t=1}^{k} \|\phi_h(s_h^t, a_h^t)\|^2_{(\Sigma_h^k)^{-1}} &= \sum_{t=1}^{k} \operatorname{tr}\left(\phi_h(s_h^t, a_h^t)^T (\Sigma_h^k)^{-1} \phi_h(s_h^t, a_h^t)\right) \\
&= \sum_{t=1}^{k} \operatorname{tr}\left((\Sigma_h^k)^{-1} \phi_h(s_h^t, a_h^t) \phi_h(s_h^t, a_h^t)^T\right) = \sum_{t=1}^{k} \frac{\lambda_i}{\lambda_i + \lambda} \leq d
\end{aligned}$$

with $\{\lambda_i\}_{i=1}^d$ being the eigenvalues of $\phi_h(s_h^t, a_h^t)\phi_h(s_h^t, a_h^t)^T$. Finally, using $\|\hat{w}_h^k\|_2 = \max_{v:\|v\|_2=1} |v^T \hat{w}_h^k|$ completes the proof. ◻

Lemma E.5 ((Jin et al., 2020))**.** *Let $\mathcal{V}(L, B, \lambda) \subset \{\mathcal{S} \to \mathbb{R}\}$ be a class of functions with the following parametric form:*

$$V(\cdot) = \min\{\max_{a \in \mathcal{A}} \phi_h(\cdot, a)^T w - \beta \|\phi(\cdot, a)\|_{\Sigma^{-1}}, H-h+1\}^+,$$

where the parameters (w, β, Σ) satisfy: $\|w\|_2 \leq L, \beta \in [0, B], \lambda_{\min}(\Sigma) \geq \lambda$. Assume $\|\phi_h(s,a)\|_2 \leq 1, \forall (s,a)$. Let N_ϵ be the ϵ-covering number of $\mathcal{V}(L, B, \lambda)$ with respect to the maximal norm $\|\cdot\|_\infty$. We have:

$$\log N_\epsilon \leq d \log(1 + 4L/\epsilon) + d^2 \log(1 + 8\sqrt{d}B^2/(\lambda\epsilon^2)).$$

E.2 Linear features

Lemma E.6. *Let $\phi : \mathcal{S} \to \mathbb{R}^d$ be any feature and p be any density with respect to the Lebesgue measure on $\mathcal{S}$. Let $A = \mathbb{E}_{s \sim p(s)}\left[\phi(s)\phi(s)^T\right]$. We have*

$$col(A) = span(\{\phi(s) : \forall s \in \mathcal{S} \ s.t.\ p(s) > 0\}),$$

where $col(A)$ denotes the column space of A.

Proof. Let $B := \text{span}(\{\phi(s) : \forall s \in \mathcal{S} \text{ s.t. } p(s) > 0\})$. We need to prove that $A = B$. First, as the i-th column of A, $\text{col}_i(A) = \int_{\{s:p(s)>0\}} \phi(s)\phi(s)_i dp(s) \in B$, where $\phi(s)_i$ denotes the i-component of $\phi(s) \in \mathbb{R}^d$. Thus, $\text{col}(A) \subseteq B$. Now we prove that $B \subseteq \text{col}(A)$.

For any $x \in \mathbb{R}^d$, we have $x^T A x = \int_{\{s:p(s)>0\}} (x^T \phi(s))^2 p(s) ds$. Thus, for any $x \in \text{null}(A)$ (i.e. $Ax = 0$), we have $x^T \phi(s) = 0, \forall s$ such that $p(s) > 0$. Hence, $\text{null}(A) \perp B$. But we have $\text{col}(A) = \text{null}(A)^\perp$, thus $B \subseteq \text{col}(A)$. □

Lemma E.7 ((Papini et al., 2021b)). *For any symmetric p.s.d. matrix $A \in \mathbb{R}^{d\times d}$ with $\|A\| > 0$, the smallest positive eigenvalue of A is:*

$$\lambda^+_{\min}(A) = \min_{x \in col(A): \|x\|_2 = 1} x^T A x.$$

Proof. Let $0 \le \lambda_1 \le \ldots \le \lambda_d$ be the eigenvalues of A with corresponding orthonormal eigenvectors $u_1, \ldots, u_d$. By the iterative representation of eigenvalues, we have:

$$\lambda_i = \min_{x \in \{u_1, \ldots, u_{i-1}\}^\perp : \|x\|_2 = 1} x^T A x.$$

Let $d' = \min\{i \in [d] : \lambda_i > 0\}$. As $\|A\| = \lambda_d > 0$, such d' exists. Thus, $\text{span}(\{u_1, \ldots, u_{d'-1}\}) = \text{null}(A)$. Note that $\text{null}(A)^T = \text{col}(A)$ as A is symmetric, we complete the proof. □

Lemma E.8. *Let $A \in \mathbb{R}^{d\times d}$ be a symmetric matrix with non-zero eigenvalues $\lambda_1 \le \lambda_2 \ldots \le \lambda_d$ and corresponding orthonormal eigenvectors $u_1, \ldots, u_d$. Assume $\|A\| > 0$ (i.e. $\lambda_d > 0$). Let $d' = \min_{i \in [d]:\lambda_i > 0}$. We have:*

$$\forall x \in col(A), \|x\|_{A^{-1}} \cdot \|x\|_A \le \frac{\lambda_d}{\lambda_{d'}}.$$

E.3 Concentration inequalities

Lemma E.9 (Hoeffding-Azuma inequality). *Suppose $\{X_k\}_{k=0}^\infty$ is a martingale, i.e. $\mathbb{E}[|X_k|] < \infty$ and $\mathbb{E}[X_{k+1}|X_k, \ldots, X_0] = X_k, \forall k$, and suppose that $\forall k$, $|X_k - X_{k-1}| \le c_k$ almost surely. Then for all positive n and t, we have:*

$$\mathbb{P}(|X_n - X_0| \ge t) \le 2\exp\left(\frac{-t^2}{\sum_{i=1}^n c_i^2}\right).$$

Lemma E.10 (A variant of Hoeffding-Azuma inequality). *Suppose $\{Z_k\}_{k=0}^\infty$ is a real-valued stochastic process with corresponding filtration $\{\mathcal{F}_k\}_{k=0}^\infty$, i.e. $\forall k$, Z_k is $\mathcal{F}_k$-measurable. Suppose that for any k, $\mathbb{E}[|Z_k|] < \infty$ and $|Z_k - \mathbb{E}[Z_k|\mathcal{F}_{k-1}]| \le c_k$ almost surely. Then for all positive n and t, we have:*

$$\mathbb{P}\left(\left|\sum_{k=1}^n Z_k - \sum_{k=1}^n \mathbb{E}[Z_k|\mathcal{F}_{k-1}]\right| \ge t\right) \le 2\exp\left(\frac{-t^2}{\sum_{i=1}^n c_i^2}\right).$$

Proof. This lemma is a direct application of Lemma E.9 with $X_k = \sum_{i=1}^k (Z_i - \mathbb{E}[Z_i|F_{i-1}])$. □

Lemma E.11 (Concentration of self-normalized processes (Abbasi-yadkori et al., 2011)). *Let $\{\eta_t\}_{t=1}^{\infty}$ be a real-valued stochastic process with corresponding filtration $\{\mathcal{F}_t\}_{t=0}^{\infty}$ (i.e. η_t is $\mathcal{F}_t$-measurable). Assume that $\eta_t|\mathcal{F}_{t-1}$ is zero-mean and R-subGaussian, i.e., $\mathbb{E}[\eta_t|\mathcal{F}_{t-1}] = 0$, and*

$$\forall \lambda \in \mathbb{R}, \mathbb{E}\left[e^{\lambda \eta_t}|\mathcal{F}_{t-1}\right] \leq e^{\lambda^2 R^2/2}.$$

Let $\{x_t\}_{t=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process where x_t is $\mathcal{F}_{t-1}$-measurable and $\|x_t\| \leq L$. Let $\Sigma_k = \lambda I_d + \sum_{t=1}^{k} x_t x_t^T$. Then for any $\delta > 0$, with probability at least $1-\delta$, it holds for all $k > 0$ that

$$\left\|\sum_{t=1}^{k} x_t \eta_t\right\|_{\Sigma_k^{-1}}^2 \leq 2R^2 \log\left[\frac{det(\Sigma_k)^{1/2} det(\Sigma_0)^{-1/2}}{\delta}\right] \leq 2R^2\left[\frac{d}{2}\log\frac{kL^2+\lambda}{\lambda} + \log\frac{1}{\delta}\right].$$

Lemma E.12 (Uniform concentration of self-normalized processes (Jin et al., 2020)). *Let $\{s_t\}_{t=1}^{\infty}$ be a stochastic process on state space $\mathcal{S}$ with corresponding filtration $\{\mathcal{F}_t\}_{t=0}^{\infty}$ (i.e. s_t is $\mathcal{F}_t$-measurable). Let $\{\phi_t\}_{t=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process where ϕ_t is $\mathcal{F}_{t-1}$-measurable and $\|\phi_t\|_2 \leq 1$. Let $\Sigma_k = \lambda I_d + \sum_{t=1}^{k-1} \phi_t \phi_t^T$. Then for any $\delta > 0$, with probability at least $1-\delta$, for all $k \geq 0$ and any $V \in \mathcal{V} \subset \{\mathcal{S} \to [0, H]\}$, we have*

$$\left\|\sum_{t=1}^{k-1} \phi_t\left(V(s_t) - \mathbb{E}\left[V(s_t)|\mathcal{F}_{t-1}\right]\right)\right\|_{\Sigma_k^{-1}}^2 \leq 4H^2\left[\frac{d}{2}\log\left(\frac{k+\lambda}{\lambda}\right) + \log\frac{N_\epsilon}{\delta}\right] + \frac{8k^2\epsilon^2}{\lambda},$$

where N_ϵ is the ϵ-covering number of $\mathcal{V}$ with respect to the distance $\|\cdot\|_\infty$.

Proof. For any $V \in \mathcal{V}$, there exists $\bar{V}$ in the ϵ-covering such that $V = \bar{V} + \Delta_V$ where $\sup_s |\Delta_V(s)| \leq \epsilon$. We have the following decomposition:

$$\begin{aligned}\left\|\sum_{t=1}^{k} \phi_t\left(V(s_t) - \mathbb{E}\left[V(s_t)|\mathcal{F}_{t-1}\right]\right)\right\|_{\Sigma_k^{-1}}^2 &\leq 2\left\|\sum_{t=1}^{k} \phi_t\left(\bar{V}(s_t) - \mathbb{E}\left[\bar{V}(s_t)|\mathcal{F}_{t-1}\right]\right)\right\|_{\Sigma_k^{-1}}^2 \\ &\quad + 2\left\|\sum_{t=1}^{k} \phi_t\left(\Delta_V(s_t) - \mathbb{E}\left[\Delta_V(s_t)|\mathcal{F}_{t-1}\right]\right)\right\|_{\Sigma_k^{-1}}^2\end{aligned}$$

where the first term can be bounded by Lemma E.11 and the second term is bounded by $8k^2\epsilon^2/\lambda$. Then using the union bound over the ϵ-covering completes the proof. □

Lemma E.13 (Freedman's inequality (Tropp, 2011)). *Let $\{X_k\}_{k=1}^{n}$ be a real-valued martingale difference sequence with the corresponding filtration $\{\mathcal{F}_k\}_{k=1}^{n}$, i.e. X_k is $\mathcal{F}_k$-measurable and $\mathbb{E}[X_k|\mathcal{F}_{k-1}] = 0$. Suppose for any k, $|X_k| \leq M$ almost surely and define $V := \sum_{k=1}^{n} \mathbb{E}\left[X_k^2|\mathcal{F}_{k-1}\right]$. For any $a, b > 0$, we have:*

$$\mathbb{P}\left(\sum_{k=1}^{n} X_k \geq a, V \leq b\right) \leq \exp\left(\frac{-a^2}{2b + 2aM/3}\right).$$

In an alternative form, for any $t > 0$, we have:

$$\mathbb{P}\left(\sum_{k=1}^{n} X_k \geq \frac{2Mt}{3} + \sqrt{2bt}, V \leq b\right) \leq e^{-t}.$$

Lemma E.14 (Matrix Freedman's inequality (Tropp, 2011))**.** *Let $\{X_k\}$ be a $d \times d$ stochastic matrices adapted to the filtration $\{F_k\}$, i.e. X_k is $\mathcal{F}_k$-measurable. Suppose that $\forall k, \| X_k - \mathbb{E}[X_k|\mathcal{F}_{k-1}] \| \leq M$ almost surely for some $M > 0$. Define the quadratic variation process*

$$V_k := \sum_{i=1}^{k} Var[X_i|\mathcal{F}_{i-1}].$$

For any $a, b \geq 0$, we have:

$$\mathbb{P}\left(\exists k \geq 0 : \| \sum_{i=1}^{k} X_i - \mathbb{E}[X_i|\mathcal{F}_{i-1}] \| \geq a, \|V_k\|_2 \leq b\right) \leq d \exp\left(\frac{-a^2}{2b + 2aM/3}\right).$$

In an alternative form, for any $t > 0$, we have:

$$\mathbb{P}\left(\exists k \geq 0 : \left\| \sum_{i=1}^{k} X_i - \mathbb{E}[X_i|\mathcal{F}_{i-1}] \right\| \geq \frac{2Mt}{3} + \sqrt{2bt}, \|V_k\|_2 \leq b\right) \leq de^{-t}.$$

F Model-based offline RL

In this section, we consider the linear mixture MDP model (Ayoub et al., 2020) that assumes that the unknown transition function is an unknown linear mixture of several basic known probabilities.

Definition 3 (Linear mixture MDP)**.** *An MDP $\mathcal{M}(\mathcal{S}, \mathcal{A}, H, \{r_h\}_{h=1}^{H} \{\mathbb{P}_h\}_{h=1}^{H})$ is said to be a linear mixture MDP if there is a known feature mapping $\phi(s'|s,a) : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$ and an unknown vector $w_h^* \in \mathbb{R}^d$ with $\|w_h^*\|_2 \leq C_w$ such that $\mathbb{P}_h(s'|s,a) = \langle \phi(s'|s,a), w_h^* \rangle$ for all (s, a, s', h) and r_h is deterministic and known (for simplicity). Moreover, for any bounded function $V : \mathcal{S} \to [0, 1]$, we have $\|\phi_V(s,a)\|_2 \leq 1$ for any (s,a), where $\phi_V(s,a) = \sum_{s' \in \mathcal{S}} \phi(s'|s,a) V(s') \in \mathbb{R}^d$.*

We consider the bootstrapped, constrained and pessimistic variant of Value-Targeted Regression (Ayoub et al., 2020) which is shown in Algorithm . The algorithm is very similar to Algorithm 1 except that we compute $\hat{w}_h^k$ by solving the following regularized least-square regression in Line 6:

$$\hat{w}_h^k \leftarrow \arg\min_{w \in \mathbb{R}^d} \lambda \|w\|_2^2 + \sum_{i=1}^{k-1} \left(\phi_{\hat{V}_{h+1}^k}(s_h^t, a_h^t)^T w - \hat{V}_{h+1}^i(s_{h+1}^i)\right)^2.$$

The flow of the results is very similar to the case of BPCPVI except some minor modifications to reflect the changes from model-free methods to model-based methods. Here we only present the results that are different from their counterpart in BPCPVI.

Lemma F.1. *In Algorithm 2, if we choose*

$$\beta_k = H \sqrt{d \log \frac{H + kH^3/\lambda}{\delta}} + \sqrt{\lambda} C_w$$

then with probability at least $1 - \delta$:

$$\forall (k, h, s, a) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}, |(T_h \hat{V}_{h+1}^k)(s,a) - (\hat{T}_h^k \hat{V}_{h+1}^k)(s,a)| \leq \beta_k \cdot \|\phi_{\hat{V}_{h+1}^k}(s,a)\|_{(\Sigma_h^k)^{-1}}.$$

Algorithm 2 Bootstrapped and Constrained Pessimistic Value-Targeted Regression (BCP-VTR)

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: $\quad\quad \Sigma_h^k \leftarrow \sum_{t=1}^{k-1} \phi_{\hat{V}_{h+1}^k}(s_h^t, a_h^t) \cdot \phi_{\hat{V}_{h+1}^k}(s_h^t, a_h^t)^T + \lambda \cdot I$.
6: $\quad\quad \hat{w}_h^k \leftarrow (\Sigma_h^k)^{-1} \sum_{t=1}^{k-1} \phi_{\hat{V}_{h+1}^k}(s_h^t, a_h^t) \cdot \hat{V}_{h+1}^k(s_{h+1}^t)$.
7: $\quad\quad b_h^k(\cdot,\cdot) \leftarrow \beta_k \cdot \|\phi_{\hat{V}_{h+1}^k}(\cdot,\cdot)\|_{(\Sigma_h^k)^{-1}}$.
8: $\quad\quad \bar{Q}_h^k(\cdot,\cdot) \leftarrow \langle \phi_{\hat{V}_{h+1}^k}(\cdot,\cdot), \hat{w}_h^k \rangle - b_h^k(\cdot,\cdot)$.
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 $\{\hat{\pi}^k : k \in [K+1]\}$.

Proof. We have:

$$(T_h \hat{V}_{h+1}^k)(s,a) = r_h(s,a) + \langle \phi_{\hat{V}_{h+1}^k}^k(s,a), w_h^* \rangle,$$
$$(\hat{T}_h \hat{V}_{h+1}^k)(s,a) = r_h(s,a) + \langle \phi_{\hat{V}_{h+1}^k}^k(s,a), \hat{w}_h^k \rangle.$$

Moreover, by (Abbasi-yadkori et al., 2011, Theorem 2), with probability at least $1-\delta$, we have:

$$\forall h \in [H], w_h^* \in \{w \in \mathbb{R}^d : \|w - \hat{w}_h^k\|_{\Sigma_h^k} \le \beta(k)\}.$$

Finally, using the inequality $\langle x, y \rangle \le \|x\|_A \cdot \|y\|_{A^{-1}}$ for any invertible matrix A and vectors x, y completes the proof. □

Theorem 7. *Under Assumption 4.1-4.2, w.p.a.l.* $1 - \Omega(\frac{1}{K})$ *over the randomness of* $\mathcal{D}$*, for the sub-optimality bound of BCP-VTR, we have:*

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

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

Theorem 8 ($\frac{\log K}{K}$-type sub-optimality bound)**.** *Under Assumption 4.1-4.2-4.3, w.p.a.l.* $1 - (1 + 3\log_2(H/\Delta_{\min}))\delta$*, for the sub-optimality bound of BCP-VTR, we have:*

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

Remark 8. *If we set the* δ *in Theorem 2 as* $\delta = \Omega(1/K)$*, then for the expected sub-optimality bound of BCP-VTR, we have:*

$$\mathbb{E}\left[\mathrm{SubOpt}(\hat{\pi}^{mix})\right] = \tilde{\mathcal{O}}\left(\frac{d^2 H^2 \kappa^3}{\Delta_{\min} \cdot K}\right).$$

Theorem 9. *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 exist a linear mixture MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r, d_0)$ 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:*

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

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

G Numerical simulation

In this appendix, we provide the details for the numerical simulation for Figure 1 in the main paper.

Linear MDP construction. We construct a simple linear MDP following (Yin et al., 2022; Min et al., 2021). We consider an MDP instance with $\mathcal{S} = \{0, 1\}$, $\mathcal{A} = \{0, 1, \cdots, 99\}$, and the feature dimension $d = 10$. Each action $a \in [99]$ is represented by its binary encoding vector $u_a \in \mathbb{R}^8$ with entry being either -1 or 1. We define

$$\delta(s, a) = \begin{cases} 1 & \text{if } \mathbb{1}\{s = 0\} = \mathbb{1}\{a = 0\}, \\ 0 & \text{otherwise.} \end{cases}$$

- The feature mapping $\phi(s, a)$ is given by

$$\phi(s, a) = [u_a^T, \delta(s, a), 1 - \delta(s, a)]^T \in \mathbb{R}^{10}.$$

- The true measure $\nu_h(s)$ is given by

$$\nu_h(s) = [0, \cdots, 0, (1 - s) \oplus \alpha_h, s \oplus \alpha_h]$$

  where $\{\alpha_h\}_{h\in[H]} \in \{0, 1\}^H$ and $\oplus$ is the XOR operator. We define

$$\theta_h = [0, \cdots, 0, r, 1 - r]^T \in \mathbb{R}^{10},$$

  where $r = 0.99$. Recall that the transition follows $P_h(s'|s, a) = \langle\phi(s, a), \nu_h(s')\rangle$ and the mean reward $r_h(s, a) = \langle\phi(s, a), \theta_h\rangle$.

Behavior policy. At state $s = 0$, choose action $a = 0$ with probability p and action $a = 1$ with probability $(1 - p)$; at state $s = 1$, choose action $a = 0$ with probability p and choose the other actions uniformly with probability $(1 - p)/99$. This behavior policy does not uniformly cover all state-space pairs but only need to satisfy Assumption 4.1.

Experiment. We computed $\mathrm{SubOpt}(\hat{\pi}^K)$ for each $K \in \{1, \ldots, 1000\}$ where $\hat{\pi}^K$ is returned by Algorithm 1. We tested for different values of $\beta \in \{0, 0.1, 0.2, 0.5, 1, 2\}$ with different episode lengths $H \in \{20, 30, 50, 80\}$. We run each experiment for 30 times and plot the mean and standard variance of the sub-optimality in Figure 1. We observe that $\beta = 1$ gives the best performance in all cases of H. It also confirms the benefit of being properly pessimistic (i.e. $\beta = 1$) versus being non-pessimistic (i.e. $\beta = 0$) for offline RL. In the case of $\beta = 1$, we observe both phenomenon in the main paper: fast rate in the first 100 episodes and zero sub-optimality in the later stage.