NEAR-OPTIMAL OFFLINE REINFORCEMENT LEARNING WITH LINEAR REPRESENTATION: LEVERAGING VARIANCE INFORMATION WITH PESSIMISM

Ming Yin∗ Department of Computer Science University of California, Santa Barbara ming.yin@cs.ucsb.edu
Yaqi Duan† Department of Electrical and Computer Engineering Princeton University yaqi@princeton.edu
Mengdi Wang† Department of Electrical and Computer Engineering Princeton University mengdi@princeton.edu
Yu-Xiang Wang∗ Department of Computer Science University of California, Santa Barbara yuxiangw@cs.ucsb.edu

ABSTRACT

Offline reinforcement learning, which seeks to utilize offline/historical data to optimize sequential decision-making strategies, has gained surging prominence in recent studies. Due to the advantage that appropriate function approximators can help mitigate the sample complexity burden in modern reinforcement learning problems, existing endeavors usually enforce powerful function representation models (e.g. neural networks) to learn the optimal policies. However, a precise understanding of the statistical limits with function representations, remains elusive, even when such a representation is linear.

Towards this goal, we study the statistical limits of offline reinforcement learning with linear model representations. To derive the tight offline learning bound, we design the variance-aware pessimistic value iteration (VAPVI), which adopts the conditional variance information of the value function for time-inhomogeneous episodic linear Markov decision processes (MDPs). VAPVI leverages estimated variances of the value functions to reweight the Bellman residuals in the least-square pessimistic value iteration and provides improved offline learning bounds over the best-known existing results (whereas the Bellman residuals are equally weighted by design). More importantly, our learning bounds are expressed in terms of system quantities, which provide natural instance-dependent characterizations that previous results are short of. We hope our results draw a clearer picture of what offline learning should look like when linear representations are provided.

1 INTRODUCTION

Offline reinforcement learning (offline RL or batch RL Lange et al. (2012); Levine et al. (2020)) is the framework for learning a reward-maximizing policy in an unknown environment (Markov Decision Process or MDP) using the logged data coming from some behavior policy π. Function approximations, on the other hand, are well-known for generalization in the standard supervised learning. Offline RL with function representation/approximation, as a result, provides generalization across large state-action spaces for the challenging sequential decision-making problems when no iteration is allowed (as opposed to online learning). This paradigm is crucial to the success of modern RL problems as many deep RL algorithms find their prototypes in the literature of offline RL. For example, Xie and Jiang (2020) provides a view that Fitted Q-Iteration (Gordon, 1999; Ernst et al., 2005) can be considered as the theoretical prototype of the deep Q-networks algorithm (DQN) Mnih et al. (2015) with neural networks being the function representors. On the empirical side, there are a huge body of deep RL-based algorithms (Mnih et al., 2015; Silver et al., 2017; Fujimoto et al., 2019; Kumar et al., 2019; Wu et al., 2019; Kidambi et al., 2020; Yu et al., 2020; Kumar et al., 2020; Janner et al., 2021; Chen et al., 2021a; Kostrikov et al., 2022) that utilize function approximations to achieve respective successes in the offline regime. However, it is also realized that practical function approximation schemes can be quite sample inefficient (e.g. millions of samples are needed for deep Q-network to solve certain Atari games Mnih et al. (2015)).

†The environment could have other forms as well, e.g. partially-observed MDP (POMDP) or non-markovian decision process (NMDP).
To understand this phenomenon, there are numerous studies consider how to achieve sample efficiency with function approximation from the theoretical side, as researchers find sample efficient algorithms are possible with particular model representations, in either online RL (e.g. Yang and Wang (2019; 2020); Modi et al. (2020); Jin et al. (2020); Ayoub et al. (2020); Jiang et al. (2017); Du et al. (2019); Sun et al. (2019); Zanette et al. (2020); Zhou et al. (2021a); Jin et al. (2021a); Du et al. (2021)) or offline RL (e.g. Munos (2003); Chen and Jiang (2019); Xie and Jiang (2020); Jin et al. (2021b); Xie et al. (2021a); Min et al. (2021); Duan et al. (2021); Nguyen-Tang et al. (2021); Zanette et al. (2021)). Among them, the linear MDP model (Yang and Wang, 2020; Jin et al., 2020), where the transition is represented as a linear combination of the given $d$-dimensional feature, is (arguably) the most studied setting in function approximation and there are plenty of extensions based upon it (e.g. generalized linear model (Wang et al., 2021b), reward-free RL (Wang et al., 2020), gap-dependent analysis (He et al., 2021) or generative adversarial learning (Liu et al., 2021)). Given its prosperity, however, there are still unknowns for understanding function representations in RL, especially in the offline case.

- While there are surging researches in showing provable sample efficiency (polynomial sample complexity is possible) under a variety of function approximation schemes, how to improve the sample efficiency for a given class of function representations remains understudied. For instance, given a neural network approximation class, an algorithm that learns the optimal policy with complexity $O(H^{10})$ is far worse than the one that can learn in $O(H^{5})$ sample complexity, despite that both algorithms are considered sample efficient. Therefore, how to achieve the optimal/tight sample complexity when function approximation is provided is a valuable question to consider. On the other hand, it is known that tight sample complexity, due to the limit of the existing statistical analysis tools, can be very tough to establish when function representation has a very complicated form. However, does this mean tight analysis is not hopeful even when the representation is linear?

- Second, in the existing analysis of offline RL (with function approximation or simply the tabular MDPs), the learning bounds depend either explicitly on the data-coverage quantities (e.g. uniform concentrability coefficients Chen and Jiang (2019); Xie and Jiang (2020), uniform visitation measure Yin et al. (2021); Yin and Wang (2021a) and single concentrability Rashidinejad et al. (2021); Xie et al. (2021b)) or the horizon length $H$ (Jin et al., 2021b; Uehara and Sun, 2021). While those results are valuable as they do not depend on the structure of the particular problem (therefore, remain valid even for pathological MDPs), in practice, the empirical performances of offline reinforcement learning are often far better than those non-adaptive bounds would indicate. Can the learning bounds reflect the nature of individual MDP instances when the MDP model has a certain function representation?

In this work, we think about offline RL from the above two aspects. In particular, we consider the fundamental linear model representations and ask the following question of interest:

**Can we achieve the statistical limits for offline RL when models have linear representations?**

### 1.1 Related Works

**Offline RL with general function representations.** The finite sample analysis of offline RL with function approximation is initially conducted by Fitted $Q$-Iteration (FQI) type algorithms and can be dated back to (Munos, 2003; Szepesvári and Munos, 2005; Antos et al., 2008a;b). Later, Chen and Jiang (2019); Le et al. (2019); Xie and Jiang (2020) follow this line of research and derive the improved learning results. However, owing to the aim for tackling general function approximation, those learning bounds are expressed in terms of the stringent concentrability coefficients (therefore, are less adaptive to individual instances) and are usually only information-theoretical, due to the computational intractability of the optimization procedure over the general function classes. Other works impose weaker assumptions (e.g. partial coverage (Liu et al., 2020; Kidambi et al., 2020; Uehara and Sun, 2021)), and their finite sample analysis are generally suboptimal in terms of $H$ or the effective horizon $\left(1 - \gamma\right)^{-1}$.

**Offline RL with tabular models.** For tabular MDPs, tight learning bounds can be achieved under several data-coverage assumptions. For the class of problems with uniform data-visitation measure $d_m$, the near-optimal sample complexity bound has the rate $O(H^3/d_m\epsilon^2)$ for time-inhomogeneous MDPs (Yin et al., 2021) and $O(H^2/d_m\epsilon)$ for time-homogeneous MDPs (Yin and Wang, 2021a; Ren...
et al., 2021). Under the single concentrability assumption, the tight rate $O(H^3SC^*/\epsilon^2)$ is obtained by Xie et al. (2021b). In particular, the recent study Yin and Wang (2021b) introduces the intrinsic offline learning bound that is not only instance-dependent but also subsumes previous optimal results.

**Offline RL with linear model representations.** Recently, there is more focus on studying the provable efficient offline RL under the linear model representations. Jin et al. (2021b) first shows offline RL with linear MDP is provably efficient by the pessimistic value iteration. Their analysis deviates from their lower bound by a factor of $d \cdot H$ (check their Theorem 4.4 and 4.6). Later, Xie et al. (2021a) considers function approximation under the Bellman-consistent assumptions, and, when realized to linear MDP setting, improves the sample complexity guarantee of Jin et al. (2021b) by an order $O(d)$ (Theorem 3.2). However, their improvement only holds for finite action space (due to the dependence $\log |A|$) and by the direct reduction (from Theorem 3.1) their result does not imply a computationally tractable algorithm with the same guarantee. Concurrently, Zanette et al. (2021) considers the Linear Bellman Complete model and designs the actor-critic style algorithm that achieves tight result under the assumption that the value function is bounded by 1. While their algorithm is efficient (which is based on solving a sequence of second-order cone programs), the resulting learning bound requires the action space to be finite due to the mirror descent updates in the Actor procedure (Agarwal et al., 2021). Besides, assuming the value function to be less than 1 simplifies the challenges in dealing with horizon $H$ since when rescaling their result to $[0,H]$, there is a $H$ factor blow-up, which makes no horizon improvement comparing to Jin et al. (2021b). As a result, none of the existing algorithms can achieve the statistical limit for the well-structured linear MDP model with the general (infinite or continuous) state-action spaces. On the other hand, Wang et al. (2021a); Zanette (2021) study the statistical hardness of offline RL with linear representations by proving the exponential lower bounds. Recently, Foster et al. (2021) shows realizability and concentrability are not sufficient for offline learning when state space is arbitrary large.

**Variance-aware studies.** Talebi and Maillard (2018) first incorporates the variance structure in online tabular MDPs and Zanette and Brunskill (2019) tightens the result. For linear mixture MDPs, Zhou et al. (2021a) first uses variance structure to achieve near-optimal result and the Weighted OFUL incorporates the variance structure explicitly in the regret bound. Recently, Variance-awareness is also considered in Zhang et al. (2021) for horizon-free setting and for OPE problem (Min et al., 2021). In particular, We point out that Min et al. (2021) is the first work that uses variance reweighting for policy evaluation in offline RL, which inspires our study for policy optimization problem. The guarantee of Min et al. (2021) strictly improves over Duan et al. (2020) for OPE problem.

1.2 **Our contribution**

In this work, we study offline RL for time-inhomogeneous episodic linear Markov decision processes. Linear MDPs serve as one critical step towards understanding function approximation in RL since: 1. unlike general function representation, linear MDP representation has the well-structured form by the given feature representors, which makes delicate statistical analysis hopeful; 2. unlike tabular representation, which only works for finite models, linear MDP provides generalization as it adapts to infinite or continuous state-action spaces. Especially, we design the variance-aware pessimistic value iteration (VAPVI, Algorithm 1) which incorporates the conditional variance information of the value function and, by the variance structure, Theorem 3.2 is able to improve over the aforementioned state-of-the-art guarantees. In addition, we further improve the state-action guarantee by designing an even tighter bonus (4). VAPVI-Improved (Theorem 3.3) is near-minimax optimal as indicated by our lower bound (Theorem 3.5). Importantly, the resulting learning bounds from VAPVI/VAPVI-Improved are able to characterize the adaptive nature of individual instances and yield different convergence rates for different problems. Algorithmically, our algorithm builds upon the nice Min et al. (2021) with pessimism as we use the estimated variances to reweight the Bellman residual learning objective so that the (training) samples with high uncertainty get less attention (Section 3). This is the key to obtaining instance-adaptive guarantees.

2 **Preliminaries**

---

2This comparison is based on translating their infinite horizon discounted setting to the finite-horizon case.
2.1 Problem settings

Episodic time-inhomogeneous linear Markov decision process. A finite-horizon Markov Decision Process (MDP) is denoted as $M = (S, A, P, r, H, d_1)$ (Sutton and Barto, 2018), where $S$ is the arbitrary state space and $A$ is the arbitrary action space which can be infinite or even continuous. A time-inhomogeneous transition kernel $P_h : S \times A \rightarrow \Delta^S$ ($\Delta^S$ represents a probability simplex) maps each state action $(s_h, a_h)$ to a probability distribution $P_h(\cdot|s_h, a_h)$ and $P_h$ can be different across time. In addition, $r : S \times A \rightarrow \mathbb{R}$ is the mean reward function satisfying $0 \leq r \leq 1$. $d_1$ is the initial state distribution. $H$ is the horizon. A policy $\pi = (\pi_1, \ldots, \pi_H)$ assigns each state $s_h \in S$ a probability distribution over actions according to the map $s_h \mapsto \pi_h(\cdot|s_h) \forall h \in [H]$ and induces a random trajectory $s_1, a_1, r_1, \ldots, s_H, a_H, r_H, s_{H+1}$ with $s_1 \sim d_1, a_h \sim \pi(\cdot|s_h), s_{h+1} \sim P_h(\cdot|s_h, a_h), \forall h \in [H]$. In particular, we adopt the linear MDP protocol from Jin et al. (2020; 2021b), meaning that the transition kernel and the mean reward function admit linear structures in the feature map.

Definition 2.1 (Linear MDPs). An episodic MDP $(S, A, H, P, r)$ is called a linear MDP with a known (unsigned) feature map $\phi : S \times A \rightarrow \mathbb{R}^d$ if there exist $d$ unknown (unsigned) measures $\nu_h = (\nu_h^{(1)}, \ldots, \nu_h^{(d)})$ over $S$ and an unknown vector $\theta_h \in \mathbb{R}^d$ such that

$$P_h(s' | s, a) = \langle \phi(s, a), \nu_h(s') \rangle, \quad r_h(s, a) = \langle \phi(s, a), \theta_h \rangle, \quad \forall s', s \in S, a \in A, h \in [H].$$

where $\|v_h(s)\|_2 \leq \sqrt{d}$ and $\max(\|\phi(s, a)\|_2, \|\theta_h\|_2) \leq 1$ for all $h \in [H]$ and $\forall a, s \in S \times A$. $\|\mu_h(S)\| = \int_S \|\mu_h(s)\| \, ds$.

V-values and Q-values. For any policy $\pi$, the V-value functions $V^\pi_h(\cdot) \in \mathbb{R}^d$ and Q-value functions $Q^\pi_h(\cdot, \cdot) \in \mathbb{R}^d \times A$ are defined as: $V^\pi_h(s) = \mathbb{E}_\pi[\sum_{t=0}^H r_t | s_0 = s]$, $Q^\pi_h(s, a) = \mathbb{E}_\pi[\sum_{t=0}^H r_t | s_0 = s, a_0 = a]$, $\forall s, a \in S$, $h \in [H]$. The performance measure is defined as $v^\pi := \mathbb{E}_{d_1} [V^\pi_1] = \mathbb{E}_{s \sim d_1} [\sum_{t=1}^H r_t]$. The Bellman (optimality) equations follow $\forall h \in [H]: Q^\pi_h = r_h + P_h V_{h+1}^\pi$, $V^\pi_h = \max_a Q^\pi_h (\cdot, a)$ (where $Q_h, V_h, P_h$ are vectors). By Definition 2.1, the Q-values also admit linear structures, i.e. $Q^\pi_h(\cdot, a)$ for some $w^\pi_h \in \mathbb{R}^d$ (Lemma H.9). Lastly, for a policy $\pi$, we denote the induced occupancy measure over the state-action space at any time $h \in [H]$ to be: for any $E \subseteq S \times A$, $d^\pi_h(E) := \mathbb{E}[(s_h, a_h) \in E | s_1 \sim d_1, a_t \sim \pi(\cdot | s_t), s_t \sim P_{t-1}(\cdot | s_{t-1}, a_{t-1}), 1 \leq t \leq h]$ and $\mathbb{E}_h[f(s, a)] := \int_{S \times A} f(s, a) d^\pi_h(s, a) \, dsda$. Here for notation simplicity we abuse $d^\pi_h(\cdot)$ to denote either probability measure or density function.

Offline learning setting. Offline RL requires the agent to learn the policy $\pi$ that maximizes $v^\pi$, provided with the historical data $D = \{(s_h, a_h, r_h, s_{h+1})\}_{h \in [H]}$ rolled out from some behavior policy $\mu$. The offline nature requires we cannot change $\mu$ and in particular we do not know the data generating distribution of $\mu$. To sum up, the agent seeks to find a policy $\pi^{alg}$ such that $v^\pi - v^{\pi^{alg}} \leq \epsilon$ for the given batch data $D$ and a given targeted accuracy $\epsilon > 0$.

2.2 Assumptions

It is known that learning a near-optimal policy from the offline data $D$ cannot be sample efficient without certain data-coverage assumptions (Wang et al., 2021a; Yin and Wang, 2021b). To begin with, we define the population covariance matrix under the behavior policy $\mu$ for all $h \in [H]$:

$$\Sigma^\mu_h := \mathbb{E}_{\mu, h} \left[ \phi(s, a) \phi(s, a)^\top \right],$$

since $\Sigma^\mu_h$ measure the coverage of state-action space for data $D$, we make the following assumption.

Assumption 2.2 (Feature Coverage). The data distributions $\mu$ satisfy the minimum eigenvalue condition: $\forall h \in [H], \delta^\mu_h := \lambda_{\min}(\Sigma^\mu_h) > 0$ and denote $\kappa = \min_h \kappa_h$. Note $\kappa$ is a system-dependent (non-universal) quantity as it is upper bounded by $1/d$ (Assumption 2 in Wang et al. (2021a)).

We make this assumption for the following reasons. First of all, our offline learning guarantee (Theorem 3.2) provides simultaneously comparison to all the policies, which is stronger than only competing with the optimal policy (whereas relaxed assumption suffices, e.g. $\sup_{x \in \mathbb{R}^d} \sum_{k=1}^d \frac{x_k^2}{\delta^\mu_k x_k} < \infty$).

---

3This definition is a standard extension over the tabular MDPs by referencing the similar notions from the bandit literature, i.e. from Multi-armed Bandit to Linear Bandit (Lattimore and Szepesvári, 2020).
As a consequence, the behavior distribution $\mu$ must be able to explore each feature dimension for the result to be valid. Second, even if Assumption 2.2 does not hold, we can always restrict our algorithmic design to the effective subspace of $\Sigma_p^M$, which causes the alternative notion of $\kappa := \min_{h \in [H]} \{ \kappa_h : s.t. \ k_h = \text{smallest positive eigenvalue at time } h \}$ (see Appendix G.1 for detailed discussions). In this scenario, learning the optimal policy cannot be guaranteed as a constant suboptimality gap needs to be suffered due to the lack of coverage and this is formed as assumption-free RL in Yin and Wang (2021b). Lastly, previous works analyzing the linear MDPs impose very similar assumptions, e.g. Xie et al. (2021a) Theorem 3.2 where $\sum_p^{-1}$ exists and Min et al. (2021) for the OPE problem.

Next, for any function $V_{h+1}(\cdot) \in [0, H-h]$, we define the conditional variance $\sigma_{V_{h+1}}^2: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}_+$ as $\sigma^2_{V_{h+1}}(s,a) := \max \{1, \text{Var}_{p_h}(V_{h+1})(s,a)\}$. Based on this definition, we can define the variance-involved population covariance matrices as: $\Lambda_{h}^p := \mathbb{E}_{\mu,h}[\sigma_{V_{h+1}}(s,a)^{-2}\phi(s,a)\phi(s,a)']$. In particular, when $V_h = V_h^*$, we use the notation $\Lambda_{h}^p$ instead.

3 ALGORITHM

Least square regression is usually considered as one of the “default” tools for handling problems with linear structures (e.g. LinUCB algorithm for linear Bandits) and finds its popularity in RL as well since Least-Square Value Iteration (LSVI, Jin et al. (2020)) is shown to be provably efficient for linear MDPs, due to that $V_{h+1}(s')$ is an unbiased estimator of $[P_h V_{h+1}](s,a)$. Concretely, it solves the ridge regression problems at each time steps (with $\lambda > 0$ being the regularization parameter):

$$\tilde{w}_h := \arg\min_{w \in \mathbb{R}^d} \sum_{k=1}^{K} \left[ \phi(s^k_h, a^k_h, w) - r^k_h - V_{h+1}(s^k_{h+1}) \right]^2$$

and has the closed-form solution $\tilde{w}_h = \Sigma^{-1}_h \sum_{k=1}^{K} \phi(s^k_h, a^k_h)[r^k_h + V_{h+1}(s^k_{h+1})]$ with $\Sigma^{-1}_h = \sum_{k=1}^{K} \phi(s^k_h, a^k_h)\phi(s^k_h, a^k_h)^\top + \lambda I$. In offline RL, this has also been leveraged in pessimistic value iteration (Jin et al., 2021b) and fitted $Q$-evaluation (Duan et al., 2020). Nevertheless, LSVI could only yield suboptimal guarantees, as illustrated by the following example.

Example 3.1. Instantiate PEVI (Theorem 4.4 in Jin et al. (2021b)) with $\phi(s, a) = \mathbf{1}_{s,a}$ (i.e. tabular MDPs)$^5$, by direct calculation the learning bound has the form $O(dH \sum_{s,a} d_h^{*}(s,a) \sqrt{\frac{1}{K \cdot \text{Var}_{p_h}(s,a)}})$ and the optimal result (Yin and Wang (2021b) Theorem 4.1) gives $O(\sum_{s,a} d_h^{*}(s,a) \sqrt{\frac{\text{Var}_{p_h}(s,a)}{K \cdot \text{Var}_{p_h}(s,a)}})$. The former has the horizon dependence $H^2$ and the latter is $H^{3/2}$ by law of total variance.

Motivation. By comparing the above two expressions, it can be seen that PEVI cannot get rid of the explicit $H$ factor due to missing the variance information (w.r.t $V^*$). If we go deeper, one could find that it might not be all that ideal to put equal weights on all the training samples in the least square objective (2), since, unlike linear regression where the randomness coming from one source distribution, we are regressing over a sequence of distributions in RL (i.e. each $s_h, a_h$ corresponds to a different distribution $P(s_h, a_h)$. and there are possibly infinite many of them). Therefore, conceptually, the sample piece $(s_h, a_h, s_{h+1})$ that has higher variance distribution $P(s_h, a_h)$ tends to be less “reliable” than the one $(s_h, a_h, s_{h+1})$ with lower variance (hence should not have equal weight in (2)). This suggests reweighting scheme might help improve the learning guarantee and reweighting over the variance of the value function stands as a natural choice.

3.1 VARIANCE-AWARE PESSIMISTIC VALUE ITERATION

Now we explain our framework that incorporates the variance information. Our design is motivated by previous Zhou et al. (2021a) (for online learning) and Min et al. (2021) (for policy evaluation). By the offline nature, we can use the independent episodic data $\mathcal{D}' = \{(s^k_h, a^k_h, r^k_h, s^k_{h+1})\}_{k \in [K]}$ (from $\mu$) to estimate the conditional variance of any $V$-values $V_{h+1}$ via the definition $[\text{Var}_h V_{h+1}](s,a) = \frac{\text{Var}_h V_{h+1}(s,a)}{\text{Var}_h V_{h+1}(s,a)}$.

$^5$This provides a valid illustration since tabular MDP is a special case of linear MDPs.
We denote quantities
\[
[P_h(V_{h+1})^2](s, a) - ([P_hV_{h+1}](s, a))^2.
\]
For the second order moment, by Definition 2.1, it holds
\[
[P_hV_{h+1}](s, a) = \int_S V_{h+1}^2(s') \, dP_h(s' \mid s, a) = \phi(s, a)^T \int_S V_{h+1}^2(s') \, d\nu_h(s').
\]
Denote \(\beta_h := \int_S V_{h+1}^2(s') \, d\nu_h(s')\), then \(P_hV_{h+1}^2 = \langle \phi, \beta_h \rangle\) and we can estimator it via:
\[
\tilde{\beta}_h = \text{argmin}_{\beta \in \mathbb{R}^d} \sum_{k=1}^{K} \left[ \phi(s_h^k, \hat{a}_h^k, \beta) - V_{h+1}(s_h^k) \right]^2 + \lambda \|\beta\|_2^2 = \sum_{k=1}^{K} \phi(s_h^k, \hat{a}_h^k) V_{h+1}(s_h^k)
\]
and, similarly, the first order moment \(P_hV_{h+1} := \langle \phi, \theta_h \rangle\) can be estimated via:
\[
\tilde{\theta}_h = \text{argmin}_{\theta \in \mathbb{R}^d} \sum_{k=1}^{K} \left[ \phi(s_h^k, \hat{a}_h^k, \theta) - V_{h+1}(s_h^k) \right]^2 + \lambda \|\theta\|_2^2 = \sum_{k=1}^{K} \phi(s_h^k, \hat{a}_h^k) V_{h+1}(s_h^k)
\]
The final estimator is defined as \(\tilde{\sigma}_{V_h}^2(\cdot, \cdot) := \max\{1, \tilde{\nu}_h V_{h+1}(\cdot, \cdot)\\} \) with \(\tilde{\nu}_h V_{h+1}(\cdot, \cdot) = \langle \phi(\cdot, \cdot), \tilde{\beta}_h \rangle_{[0, (H-h+1)^2]} - \langle \phi(\cdot, \cdot), \tilde{\theta}_h \rangle_{[0, (H-h+1)^2]}\). In particular, when setting \(V_{h+1} = \tilde{V}_{h+1}\), it recovers \(\hat{\sigma}_h\) in Algorithm 1 line 8. Here \(\hat{\Sigma}_h = \sum_{\tau=1}^{K} \phi(s_{h}^{\tau}, \hat{a}_{h}^{\tau}) \phi(s_{h}^{\tau}, \hat{a}_{h}^{\tau})^T + \lambda I_d\).

**Variance-weighted LSVI.** The idea of LSVI (2) is based on approximate the Bellman updates:
\(T_h(V)(s, a) = r_h(s, a) + \langle P_hV(s, a)\rangle\). With variance estimator \(\tilde{\sigma}_h\) at hand, we can modify (2) to solve the variance-weighted LSVI instead (Line 10 of Algorithm 1)
\[
\tilde{w}_h := \text{argmin}_{w \in \mathbb{R}^d} \lambda \|w\|_2^2 + \sum_{k=1}^{K} \left[ \phi(s_h^k, a_h^k, w) - r_h^k - \tilde{V}_h(s_h^k) \right]^2 + \tilde{\sigma}_h^{-1} \sum_{k=1}^{K} \phi(s_h^k, a_h^k) (r_h^k + \tilde{V}_h(s_h^k))
\]
where \(\tilde{\lambda}_h = \sum_{k=1}^{K} \phi(s_h^k, a_h^k) \phi(s_h^k, a_h^k)^T / \tilde{\sigma}_h^{-1} + \lambda I_d\). The estimated Bellman update \(\tilde{T}_h\) (acts on \(\tilde{V}_{h+1}\)) is defined as: \(\tilde{T}_h(\tilde{V}_{h+1})(\cdot, \cdot) = \phi(\cdot, \cdot)^T \tilde{w}_h\) and the pessimism \(\Gamma_{h+1}\) is assigned to update \(\tilde{Q}_h \approx \tilde{T}_h \tilde{V}_{h+1} - \Gamma_h\), i.e. Bellman update + Pessimism (Line 10-12 in Algorithm 1).

**Tighter Pessimistic Design.** To improve the learning guarantee, we create a tighter penalty design that includes \(\widetilde{\lambda}_h^{-1}\) rather than \(\hat{\Sigma}_h^{-1}\) and an extra higher order \(O(\frac{1}{K})\) term:
\[
\Gamma_h \leftarrow O \left( \sqrt{d} \cdot (\phi(\cdot, \cdot)^T \tilde{\lambda}_h^{-1} \phi(\cdot, \cdot))^{1/2} + \frac{2H^3 \sqrt{d}}{K} \right)
\]
Note such a design admits no explicit factor in \(H\) in the main term (as opposed to Jin et al. (2021b)) thereby is the key for achieving adaptive/problem-dependent results (as we shall discuss later).

The full algorithm VAPVI is stated in Algorithm 1. In particular, we halve the offline data into two independent parts with \(D = \{(s_{h}^k, a_{h}^k, r_{h}^k, s_{h+1}^k)\}_{h \in [H]}\) and \(D' = \{(s_{h}^{\tau}, \hat{a}_{h}^{\tau}, \tilde{r}_{h}^{\tau}, s_{h+1}^{\tau})\}_{\tau \in [K]}\) for different purposes (estimating variance and updating \(Q\)-values).

### 3.2 Main Result

We denote quantities \(\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4\) as in the notation list A. Then VAPVI provides the following result. The complete proof is provided in Appendix C.

**Theorem 3.2.** Let \(K\) be the number of episodes. If \(K > \max\{\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4\}\) and \(\sqrt{d} > \xi\), where \(\xi := \sup_{V \in [0, H]} \sup_{s' \sim P_h(s, a), h \in [H]} \| r_{h+1} + V(s') - (T_hV)(s, a) \|_{\sigma_V(s, a)} \).

Then for any \(0 < \lambda < \kappa\), with probability \(1 - \delta\), for all policy \(\pi \) simultaneously, the output \(\hat{\pi}\) of Algorithm 1 satisfies
\[
v^\pi - v^\hat{\pi} \leq \tilde{O}(\sqrt{d} \cdot H \sum_{h=1}^{H} \mathbb{E}_\pi \left[ \sqrt{\phi(\cdot, \cdot)^T \lambda_h^{-1} \phi(\cdot, \cdot)} \right] + \frac{2H^3 \sqrt{d}}{K}
\]
where \(\lambda_h = \sum_{k=1}^{K} \phi(s_h^k, a_h^k) \phi(s_h^k, a_h^k)^T + \lambda I_d\). In particular, we have with probability \(1 - \delta\),
\[
v^\pi - v^\hat{\pi} \leq \tilde{O}(\sqrt{d} \cdot H \sum_{h=1}^{H} \mathbb{E}_\pi \left[ \sqrt{\phi(\cdot, \cdot)^T \lambda_h^{-1} \phi(\cdot, \cdot)} \right] + \frac{2H^3 \sqrt{d}}{K}
\]
where \(\lambda_h^* = \sum_{k=1}^{K} \phi(s_h^k, a_h^k) \phi(s_h^k, a_h^k)^T + \lambda I_d\) and \(\tilde{O}\) hides universal constants and the Polylog terms.

---

\(\text{The truncation used here is a standard treatment for making the estimator to be within the valid range.}\)
Algorithm 1 Variance-Aware Pessimistic Value Iteration (VAPVI)

1: Input: Dataset $\mathcal{D} = \{(s_h^t, a_h^t, r_h^t)\}_{r,h=1}^{K,H}$. Universal constant $C$.
2: Initialization: Set $\hat{V}_{h+1} = 0$.
3: for $h = H, H-1, \ldots, 1$ do
4:   $\diamond$ Phase 1: Regular Least-square Value Iteration for conditional variances
5:   Set $\hat{\Sigma}_h \leftarrow \sum_{r=1}^{K} \phi(s_h^t, a_h^t) \phi(s_h^t, a_h^t)^\top + \lambda I$
6:   Set $\hat{\beta}_h \leftarrow \sum_{r=1}^{K} \phi(s_h^t, a_h^t)^\top \hat{V}_{h+1} \cdot (s_h^t)^2$
7:   Set $\hat{\theta}_h \leftarrow \sum_{r=1}^{K} \phi(s_h^t, a_h^t)^\top \hat{V}_{h+1} (s_h^t + 1)$
8:   Set $\hat{\mathbb{V}}_{\mathbb{V},h+1}^{2} = \langle \phi(\cdot, \cdot)^\top \hat{\Lambda}_h^{-1} \phi(\cdot, \cdot) \rangle^{1/2}$
9:   Set $\hat{\sigma}_h^{2} \leftarrow \max\{1, \mathbb{V}_{\mathbb{V},h+1}^{2}(\cdot, \cdot)\}$
10: $\diamond$ Phase 2: Weighted Least-square Value Iteration for pessimistic updates
11: Set $\hat{\Lambda}_h \leftarrow \sum_{r=1}^{K} \phi(s_h^t, a_h^t) \phi(s_h^t, a_h^t)^\top / \hat{\sigma}_h^{2}(s_h^t, a_h^t) + \lambda \cdot I$
12: Set $\hat{\omega}_h \leftarrow \hat{\Lambda}_h^{-1} \cdot (\sum_{r=1}^{K} \phi(s_h^t, a_h^t) \cdot (r_h^t + \hat{V}_{h+1} (s_h^t + 1))) / \hat{\sigma}_h^{2}(s_h^t, a_h^t)$
13: Set $\Gamma_h(\cdot, \cdot) \leftarrow C \sqrt{\hat{d}} \cdot \langle \phi(\cdot, \cdot)^\top \hat{\Lambda}_h^{-1} \phi(\cdot, \cdot) \rangle^{1/2} + \frac{2H^3 \sqrt{\sigma}}{K}$. (Use $\Gamma_h$ for the improved version)
14: Set $\hat{Q}_h(\cdot, \cdot) \leftarrow \langle \phi(\cdot, \cdot)^\top \hat{\omega}_h - \Gamma_h(\cdot, \cdot) \rangle$
15: Set $\hat{Q}_h(\cdot, \cdot) \leftarrow \min\{\hat{Q}_h(\cdot, \cdot), H - h + 1\}$
16: Set $\hat{\pi}_h(\cdot | \cdot) \leftarrow \arg\max_{\pi} \langle \hat{Q}_h(\cdot, \cdot), \pi(\cdot | \cdot) \rangle \cdot A$, $\hat{V}_h(\cdot) \leftarrow \max_{\pi} \langle \hat{Q}_h(\cdot, \cdot), \pi(\cdot | \cdot) \rangle \cdot A$
17: end for
18: Output: $\{\hat{\pi}_h\}_{h=1}^{H}$.

Theorem 3.2 provides improvements over the existing best-known results and we now explain it. However, before that, we first discuss about our theorem condition.

Comparing to Zhou et al. (2021a). In the online regime, Zhou et al. (2021a) is the first result that achieves optimal regret rate with $O(dH\sqrt{T})$ in the linear (mixture) MDPs. However, this result requires the condition $d \geq H$ (their Theorem 6 and Remark 7). In offline RL, VAPVI only requires a milder condition $\sqrt{d} > \xi$ comparing to $d \geq H$ (since for any fixed $V \in [0, H]$, the standardized quantity $\frac{r + V(s') - (\hat{f}_h V)(s, a)}{\sigma_f (s, a)}$ is bounded by constant with high probability, e.g. by chebyshev inequality), which makes our result apply to a wider range of linear MDPs.

Comparing to Jin et al. (2021b). Jin et al. (2021b) first shows pessimistic value iteration (PEVI) is provably efficient for Linear MDPs in offline RL. VAPVI improves PEVI over $O(\sqrt{d})$ on the feature dimension, and improves the horizon dependence as $\Lambda_h \gg \frac{\sqrt{d}}{H} \sum_h \Lambda_h$ implies $\Lambda_h^{-1} \approx H^2 \Sigma_h^{-1}$. In addition, when instantiate to the tabular case, i.e. $\phi(s, a) = 1_{s,a}$, VAPVI gives $O(\sqrt{d} \sum_{s,a} d_h^s(s,a))$, which enjoys $O(\sqrt{H})$ improvement over PEVI (recall Example 3.1) and the order $O(H^{3/2})$ is tight (check Section G for the detailed derivation).

Comparing to Xie et al. (2021a). Their linear MDP guarantee in Theorem 3.2. enjoys the same rate as VAPVI in feature dimension but the horizon dependence is essentially the same as Jin et al. (2021b) (by translating $H \approx O(\sqrt{d} \gamma)$) therefore is not optimal. The general function approximation scheme in Xie et al. (2021a) provides elegant characterizations for on-support error and off-support error, but the algorithmic framework is information-theoretical only (and the practical version PSPI will not yield the same learning guarantee). Also, due to the use finite function class and policy class, the reduction to linear MDP only works with finite action space. As a comparison, VAPVI has no constraints on any of these.

Comparing to Zanette et al. (2021). Concurrently, Zanette et al. (2021) considers offline RL with the linear Bellman complete model, which is more general than linear MDPs and, with the assumption $Q^* \leq 1$, their PACLE algorithm provides near-minimax optimal guarantee in this setting. However, when recovering to the standard setting $Q^* \in [0, H]$, their bound will rescale by an $H$ factor, which could be suboptimal due to the variance-unawareness. The reason behind this is: when $Q^* \leq 1$, lack of variance information encoding will not matter, since in this case $\mathbb{V}_P(V^\pi) \leq 1$ has constant

---

Footnote 2 on Page 9.
order (therefore will not affect the optimal rate); when \( Q^\pi \in [0, H] \), \( \text{Var}_\pi(V^\pi) \) can be as large as \( H^3 \), effectively leveraging the variance information can help improve the sample efficiency, e.g., via law of total variances, just like VAPVI does. On the other hand, their guarantee also requires finite action space, due to the mirror descent style analysis. Nevertheless, we do point out Zanette et al. (2021) has improved state-action measure than VAPVI, as \( \| E_\pi[\phi(\cdot,\cdot)] \|_{M^{-1}} \leq \| E_\pi[\phi(\cdot,\cdot)] \|_{M^{-1}} \) by Jensen’s inequality and that norm \( \| \cdot \|_{M^{-1}} \) is convex for some positive-definite matrix \( M \).

**Adaptive characterization and faster convergence.** Comparing to existing works, one major improvement is that the main term for VAPVI \( \sqrt{d} \sum_{h=1}^H \mathbb{E}_\pi \left[ \sqrt{\phi(\cdot,\cdot)^T \Lambda_h^{-1} \phi(\cdot,\cdot)} \right] \) admits no explicit dependence on \( H \), which provides a more adaptive/instance-dependent characterization. For instance, if we ignore the technical treatment by taking \( \lambda = 0 \) and \( \sigma_h^2 \approx \text{Var}_\pi(V^\pi_h) \), then for the partially deterministic systems (where there are \( t \) stochastic \( P_h \)'s and \( H - t \) deterministic \( P_h \)'s), the main term diminishes to \( \sqrt{d} \sum_{h=1}^t \mathbb{E}_\pi \left[ \sqrt{\phi(\cdot,\cdot)^T \Lambda_h^{-1} \phi(\cdot,\cdot)} \right] \) with \( h \in \{ h : \text{s.t.} P_h \text{ is stochastic} \} \) and can be a much smaller quantity when \( t \ll H \). Furthermore, for the fully deterministic system, VAPVI automatically provides faster convergence rate \( O\left( \frac{1}{K} \right) \) from the higher order term, given that the main term degenerates to 0. Those adaptive/instance-dependent features are not enjoyed by (Xie et al., 2021a; Zanette et al., 2021), as they always provide the standard statistical rate \( O\left( \frac{1}{\sqrt{K}} \right) \) (also check Remark C.9 for a related discussion).

### 3.3 VAPVI-I: Further improvement in state-action dimension

Can we further improve the VAPVI? Indeed, by deploying a carefully tuned tighter penalty, we are able to further improve the state-action dependence if the feature is non-negative (\( \phi \geq 0 \)). Concretely, we replace the following \( \Gamma_h \) in Algorithm 1 instead, and call the algorithm VAPVI-I (or VAPVI-I for short). The proof can be found in Appendix D.

\[
\Gamma_h(s, a) \leftarrow \phi(s, a)^T \left[ \Lambda_h^{-1} \sum_{\tau=1}^K \phi(s_{\tau}, a_{\tau}) \cdot \left( r_{\tau} + \hat{V}_{h+1}(s_{\tau+1}) - \left( \hat{\tau}_h \hat{V}_{h+1}(\cdot) \right)(s_{\tau}, a_{\tau}) \right) \right] + \tilde{O}\left( \frac{H^3d}{K} \right) \quad (4)
\]

**Theorem 3.3.** Suppose the feature is non-negative (\( \phi \geq 0 \)). Let \( K \) be the number of episodes. If \( K > \max\{ M_1, M_2, M_3, M_4 \} \) and \( \sqrt{d} \gg \xi \). Deploying \( \Gamma_h \) (4) in Algorithm 1. Then for any \( 0 < \lambda < k \), with probability \( 1 - \delta \), for all policy \( \pi \) simultaneously, the output \( \hat{\pi} \) of Algorithm 1 (VAPVI-I) satisfies

\[
v^\pi - v^\hat{\pi} \leq \tilde{O}\left( \sqrt{d} \cdot \sum_{h=1}^H \mathbb{E}_\pi[\phi(\cdot,\cdot)^T \Lambda_h^{-1} \phi(\cdot,\cdot)] \right) + \tilde{O}\left( \frac{H^3d}{K} \right)
\]

In particular, when choosing \( \pi = \pi^* \), the above guarantee holds true with \( \Lambda_h^{-1} \) replaced by \( \Lambda_h^{-1} \). Here \( \Lambda_h^{-1}, \Lambda_h^{-1}, \xi \) are defined the same as Theorem 3.2.

Theorem 3.3 maintains nearly all the features of Theorem 3.2 (except higher order term is slightly worse) and the dominate term evolves from \( \mathbb{E}_\pi[\phi(\cdot,\cdot)^T \Lambda_h^{-1} \phi(\cdot,\cdot)] \) to \( \mathbb{E}_\pi[\phi(\cdot,\cdot)] \). Clearly, the two bounds differ by the magnitude of Jensen’s inequality. To provide a concrete view of how much improvement is made, we check the parameter dependence in the context of tabular MDPs (where we ignore the higher order term for conciseness). In particular, we compare the results under the single-policy concentrability.

**Assumption 3.4** (Rashidinejad et al., 2021a; Xie et al., 2021b). There exists an optimal policy \( \pi^* \), s.t. \( \sup_{s,a} d_h^{\pi^*}(s,a) =: C^* < \infty \), where \( d^\pi \) is the marginal state-action probability under \( \pi \).

In tabular RL, \( \phi(s,a) = 1_{s,a} \) and \( d = S \cdot A \) (\( S, A \) be the finite state, action cardinality), then

\[
\text{Theorem 3.2} \rightarrow \sqrt{SA} \sum_{h}^{H} \sum_{s,a} d_{h}^{\pi^*}(s,a) \sqrt{\text{Var}_{\pi,s,a}(r+V_{h+1})} \leq \sqrt{\frac{H^3C^*SA}{K}}.
\]

\[
\text{Theorem 3.3} \rightarrow \sqrt{SA} \sum_{h}^{H} \sum_{s,a} d_{h}^{\pi^*}(s,a)^2 \frac{\text{Var}_{\pi,s,a}(r+V_{h+1})}{K \cdot d_{h}^{\pi^*}(s,a)} \leq \sqrt{\frac{H^3C^*SA}{K}}.
\]

(5)
Theorem 3.3 enjoys a $S$ state improvement over Theorem 3.2 and nearly recovers the minimax rate $\sqrt{H^dC^sS/K}$ (Xie et al., 2021b). The detailed derivation can be found in Appendix G. Also, to show our result is near-optimal, we provide the corresponding lower bound. The proof is in Appendix E.

**Theorem 3.5** (Minimax lower bound). There exist a pair of universal constants $c, c' > 0$ such that given dimension $d$, horizon $H$ and sample size $K > c'd^3$, one can always find a family of linear MDP instances $\mathcal{M}$ such that (where $\Lambda^*_h = \sum_{k=1}^K \phi(s^*_h, a^*_h)^\top \phi(s^*_h, a^*_h)^\top_{1/h} V_{a_h}^k(s_h^*, a_h^*)^{-1}$ satisfies $(\Lambda^*_h)^{-1}$ exists and $\text{Var}_{h}(V^*_h(s_h^k, a_h^k)) > 0 \forall M \in \mathcal{M}$

\[
\inf_{M} \sup_{M \in \mathcal{M}} \mathbb{E}_M [v^* - v^\pi] / \left(\sqrt{d} \sum_{h=1}^H \sqrt{\mathbb{E}_\pi[v^\pi]} (\Lambda^*_h)^{-1} \mathbb{E}_\pi[v^\pi] \geq c.
\]

(6)

Theorem 3.5 nearly matches the main term in VAPVI-I (Theorem 3.3) and certifies it is near-optimal.

## 4 Proof Overview

Due to the space constraint, we could only provide a brief overview of the key proving ideas of the theorems. We begin with Theorem 3.2. First, by the extended value difference lemma (Lemma H.7), we can convert bounding the suboptimality gap of $v^* - v^\pi$ to bounding $\sum_{h=1}^H 2 \cdot \mathbb{E}_\pi [\Gamma_h(s_h, a_h)]$, given that $|T_h V_{a_h+1} - T_h V_{a_h+1}(s, a)| \leq \Gamma_h(s, a)$ for all $s, a, h$. To bound $T_h V_{a_h+1} - T_h V_{a_h+1}$, by decomposing it reduces to bounding the key quantity

\[
\phi(s, a) \sum_{r=1}^K \phi(s_h^r, a_h^r) \cdot \frac{r_h + V_{\hat{h}+1}^\top (s_h^r, a_h^r) - (T_h V_{\hat{h}+1}^\top (s_h^r, a_h^r))}{\sigma_h^2(s_h^r, a_h^r)}
\]

(7)

The term is treated in two steps. First, we bound the gap of $\left|\sigma_h^2(s_h^r, a_h^r) - \sigma^2_h\right|$ so we can convert $\sigma^2_h$ to $\sigma^2_{\hat{h}}$. Next, since $\text{Var} \left[\frac{r_h + V_{\hat{h}+1}^\top (s_h^r, a_h^r) - (T_h V_{\hat{h}+1}^\top (s_h^r, a_h^r))}{\sigma_h^2(s_h^r, a_h^r)} \right] \approx \sigma_h^2$, therefore by the variance-weighted scheme in (equation 7), we can leverage the recent technical development Bernstein inequality for self-normalized martingale (Lemma H.3) for acquiring the tight result, in contrast to the previous treatment of Hoeffding inequality for self-normalized martingale + Covering.

For the second part, one needs to further convert $\sigma^2_{\hat{h}}$ to $\sigma^2_h$ (Lemma H.3) with appropriate concentrations. The proof of Theorem 3.3 is similar but with more complicated computations and relies on using the linear representation of $\phi$ in $\Gamma_h^\top$ (4), so that the expectation over $\pi$ is inside the square root by taking expectation over the linear representation at the beginning. The lower bound proof uses a simple modification of (Zanette et al., 2021) which consists of the reduction from learning to testing with Assouad’s method, and the use of standard information inequalities (e.g. from total variation to KL divergence). For completeness, we provide the full proof in Appendix E.

## 5 Discussion and Conclusion

This work studies offline RL with linear MDP representation and contributes Variance Aware Pessimistic Value Iteration (VAPVI) which adopts the conditional variance information of the value function. VAPVI uses the estimated variances to reweight the Bellman residuals in the least-square pessimistic value iteration and provides improved offline learning bounds over the existing best-known results. VAPVI-I further improves over VAPVI in the state-action dimension and is near-minimax optimal. One highlight of the theorems is that our learning bounds are expressed in terms of system quantities, which automatically provide natural instance-dependent characterizations that previous results are short of.

On the other hand, while VAPVI/VAPVI-I close the existing gap from previous literature (Jin et al., 2021b; Xie et al., 2021a), the optimal guarantee is in the minimax sense. Although our upper bounds possess instance-dependent characterizations, the lower bound only holds true for a class of hard instances. In this sense, whether “instance-dependent optimality” can be achieved remains elusive in the current linear MDP setting (such a discussion is recently initiated in MAB problems (Xiao et al., 2021)). We leave this as future work.
ACKNOWLEDGMENTS

The authors would like to thank Quanquan Gu for explaining Min et al. (2021) and introducing a couple of related literatures. Ming Yin would like to thank Zhuoran Yang for the helpful suggestions and Dan Qiao for a careful proofreading. Mengdi Wang gratefully acknowledges funding from Office of Naval Research (ONR) N00014-21-1-2288, Air Force Office of Scientific Research (AFOSR) FA9550-19-1-0203, and NSF 19-589, CMMI-1653435. Yu-Xiang Wang gratefully acknowledges funding from National Science Foundation (NSF) #2007117 and #2003257.

REFERENCES


Appendix

A Notation List

\[
\begin{align*}
\Sigma_h & \quad \mathbb{E}_{\mu,h} \left[ \phi(s,a)\phi(s,a)^\top \right] \\
\Lambda_h & \quad \mathbb{E}_{\mu,h} \left[ \sigma_{V_{h+1}} (s,a)^{-2} \phi(s,a)\phi(s,a)^\top \right] \\
\kappa & \quad \min_h \lambda_{\min}(\Sigma_h) \\
\iota & \quad \min_h \lambda_{\min}(\Lambda_h) \geq \kappa/H^2 \text{ for any } V_h \\
\sigma^2_V(s,a) & \quad \max\{1, \text{Var}_{P_h}(V)(s,a)\} \text{ for any } V \\
\sigma^2_{V_{h+1}}(s,a) & \quad \max\{1, \text{Var}_{P_h}(V_{h+1})(s,a)\} \\
\hat{\sigma}^2_{V_h}(s,a) & \quad \max\{1, \text{Var}_{P_h}V_{h+1}(s,a)\} \\
\hat{\Sigma}_h & \quad \sum_{\tau=1}^K \phi(\tilde{s}^\tau_h, \tilde{a}^\tau_h)\phi(\tilde{s}^\tau_h, \tilde{a}^\tau_h)^\top + \lambda I_d \\
\hat{\Lambda}_h & \quad \sum_{k=1}^K \phi(s^k_h, a^k_h)\phi(s^k_h, a^k_h)^\top / \hat{\sigma}^2(s^k_h, a^k_h) + \lambda I_d \\
M_1 & \quad \max\{2\lambda, 128 \log(2d/\delta), 128H^4 \log(2d/\delta)/\kappa^2 \} \\
M_2 & \quad \max\{\kappa \log((\lambda + K)/H/\lambda_0), 962H^2 d \log((\lambda + K)H/\lambda_0)/\kappa^5 \} \\
M_3 & \quad \max\{512H^4/\kappa^2 \log(2d/\delta), 4\lambda H^2/\kappa \} \\
M_4 & \quad 12\sqrt{H^4 d \log((\lambda + K)H/\lambda_0)/\kappa} \\
\delta & \quad \text{Failure probability} \\
\xi & \quad \sup_{V \in [0,H], s' \sim P_h(s,a), h \in [H]} \left| \frac{r_h + V(s') - (T_h V)(s,a)}{\sigma_V(s,a)} \right| \\
C_{H,d,\kappa,K} & \quad 36 \sqrt{H^d d \log \left( \frac{\kappa H^d d H^2}{\lambda_0} \right)} + 12\lambda H^2 \sqrt{d} 
\end{align*}
\]

B Extended Literature Review

B.1 Linear model representation and its extension in online RL

There are numerous works in online RL that study linear model representations. Yang and Wang (2019; 2020); Jin et al. (2020) propose Linear MDP, which assumes the transition kernel and the reward are linear in given features. Cai et al. (2020); Ayoub et al. (2020); Modi et al. (2020); Zhou et al. (2021b) propose Linear mixture MDP, which assumes the transition probability is a linear combination of some base kernels. Linear Bellman Complete model (Zanette et al., 2020) generalizes linear MDP model by allowing linear functions to approximate the Q-function and the function class is closed under the Bellman update. The notion of low Bellman rank (Jiang et al., 2017) subsumes not only linear MDPs but also models including linear quadratic regulator (LQR), Reactive POMDP (Krishnamurthy et al., 2016) and Block MDP (Du et al., 2019). There are also other models, e.g. factored MDP (Sun et al., 2019), Bellman Eluder dimension (Jin et al., 2021a) and Bilinear class (Du et al., 2021). With the linear MDP model itself, there are also fruitful extensions, e.g. gap-dependent analysis with logarithmic regret (He et al., 2021), low-switching cost RL (Gao et al., 2021), safe RL (Amani et al., 2021), reward-free RL (Wang et al., 2020), generalized linear model (GLM) (Wang et al., 2021b), two-player markov game (Chen et al., 2021b) and generative adversarial learning (Liu et al., 2021). In particular, Zhou et al. (2021a) shows UCRL-VTR+ is near-minimax optimal when feature dimension \(d \geq H\).

B.2 Existing results in offline RL with model representations

Offline RL with general function representations. The finite sample analysis of offline RL with function approximation is initially conducted by Fitted Q-Iteration (FQI) type algorithms and can be dated back to (Munos, 2003; Szepesvári and Munos, 2005; Antos et al., 2008a:b). Later, Chen and Jiang (2019); Le et al. (2019); Xie and Jiang (2020) follow this line of research and derive the
improved learning results. However, owing to the aim for tackling general function approximation, those learning bounds are expressed in terms of the stringent concentrability coefficients (therefore, are less adaptive to individual instances) and are usually only information-theoretical, due to the computational intractability of the optimization procedure over the general function classes. Other works impose weaker assumptions (e.g. partial coverage (Liu et al., 2020; Kidambi et al., 2020; Uehara and Sun, 2021)), and their finite sample analysis are generally suboptimal in terms of $H$ or the effective horizon $(1 - \gamma)^{-1}$.

### Offline RL with tabular models.

For tabular MDPs, tight learning bounds can be achieved under several data-coverage assumptions. For the class of problems with uniform data-visititation measure $d_m$, the near-optimal sample complexity bound has the rate $O(H^3/d_m\epsilon^2)$ for time-inhomogeneous MDPs (Yin et al., 2021) and $O(H^2/d_m\epsilon)$ for time-homogeneous MDPs (Yin and Wang, 2021a; Ren et al., 2021). Under the single concentrability assumption, the tight rate $O(H^3SC^*/\epsilon^2)$ is obtained by Xie et al. (2021b). In particular, the recent study Yin and Wang (2022) introduces the intrinsic offline learning bound that is not only instance-dependent but also subsumes previous optimal results. More recently, Shi et al. (2022) uses model-free approach to achieve minimax rate with a larger $\epsilon$-range.

### Offline RL with linear model representations.

Recently, there are more focus on studying the provable efficient offline RL under the linear model representations. Jin et al. (2021b) first shows offline RL with linear MDP is provably efficient by the pessimistic value iteration (PEVI), which is an offline counterpart of LSVI-UCB in Jin et al. (2020). Their analysis deviates from their lower bound by a factor of $d \cdot H$ (check their Theorem 4.4 and 4.6). Later, Xie et al. (2021a) considers function approximation under the Bellman-consistent assumptions, and, when realized to linear MDP setting, improve the sample complexity guarantee of Jin et al. (2021b) by a order $O(d)$ (Theorem 3.2). However, their improvement only holds for finite action space (due to the dependence $\log |\mathcal{A}|$) and by the direct application of Lemma H.8 and recall $\Gamma_h(s,a) = \arg \max_{\pi_h} \langle \hat{Q}_h(s,\cdot), \pi_h(\cdot|s) \rangle_{\mathcal{A}}$, then by the direct application of Lemma H.8

$$V^\pi_1(s) - V^\pi_2(s) \leq \sum_{h=1}^H \mathbb{E}_\pi [\zeta_h(s_h, a_h) \mid s_1 = s] - \sum_{h=1}^H \mathbb{E}_\pi [\zeta_h(s_h, a_h) \mid s_1 = s].$$

The next lemma shows it is sufficient to bound the pessimistic penalty, which is the key in the proof.

**Lemma C.1.** Suppose with probability $1 - \delta$, it holds for all $h, s, a \in [H] \times \mathcal{S} \times \mathcal{A}$ that $|\langle \Gamma_h \hat{V}_{h+1} - \Gamma_h \hat{V}_{h+1} \rangle_\mathcal{A} | \leq \Gamma_h(s, a)$, then it implies $\forall s, a, h \in \mathcal{S} \times \mathcal{A} \times [H], 0 \leq \zeta_h(s, a) \leq 2\Gamma_h(s, a)$.  

### Proofs in Section 3.2

Instead of proofing the result for $v^* - v^{\hat{\pi}}$, in most parts of the proof we deal with $V^\pi_1 - V^\pi_2$, which is more general.

#### C.1 Some preparations

Define the Bellman update error $\zeta_h(s,a) := \langle \Gamma_h \hat{V}_{h+1} \rangle_\mathcal{A} - \hat{Q}_h(s,a)$ and recall $\hat{\pi}_h(s) = \arg \max_{\pi_h} \langle \hat{Q}_h(s,\cdot), \pi_h(\cdot|s) \rangle_{\mathcal{A}}$, then by the direct application of Lemma H.8

$$V^\pi_1(s) - V^\pi_2(s) \leq \sum_{h=1}^H \mathbb{E}_\pi [\zeta_h(s_h, a_h) \mid s_1 = s] - \sum_{h=1}^H \mathbb{E}_\pi [\zeta_h(s_h, a_h) \mid s_1 = s].$$

The next lemma shows it is sufficient to bound the pessimistic penalty, which is the key in the proof.

**Lemma C.1.** Suppose with probability $1 - \delta$, it holds for all $h, s, a \in [H] \times \mathcal{S} \times \mathcal{A}$ that $|\langle \Gamma_h \hat{V}_{h+1} - \Gamma_h \hat{V}_{h+1} \rangle_\mathcal{A} | \leq \Gamma_h(s, a)$, then it implies $\forall s, a, h \in \mathcal{S} \times \mathcal{A} \times [H], 0 \leq \zeta_h(s, a) \leq 2\Gamma_h(s, a)$.
Furthermore, it holds for any policy $\pi$ simultaneously, with probability $1 - \delta$,

$$V^\pi_1(s) - V^\pi_1(s) \leq \sum_{h=1}^{H} 2 \cdot E_\pi [\Gamma_h(s_h, a_h) \mid s_1 = s].$$

Proof of Lemma C.1. We first show given $|[(T_h \hat{V}_{h+1} - \hat{T}_h \hat{V}_{h+1})(s, a)]| \leq \Gamma_h(s, a)$, then $0 \leq \zeta_h(s, a) \leq 2\Gamma_h(s, a), \forall s, a, h \in S \times A \times [H].$

**Step 1:** we first show $0 \leq \zeta_h(s, a), \forall s, a, h \in S \times A \times [H].$

Indeed, if $\hat{Q}_h(s, a) \leq 0$, then by definition $\hat{Q}_h(s, a) = 0$ and in this case $\zeta_h(s, a) := (T_h \hat{V}_{h+1})(s, a) - \hat{Q}_h(s, a) = (T_h \hat{V}_{h+1})(s, a) \geq 0$; if $\hat{Q}_h(s, a) > 0$, then $\hat{Q}_h(s, a) \leq Q_h(s, a)$ and

$$\zeta_h(s, a) := (T_h \hat{V}_{h+1})(s, a) - \hat{Q}_h(s, a) \geq (T_h \hat{V}_{h+1})(s, a) - Q_h(s, a)$$

$$= (T_h \hat{V}_{h+1})(s, a) - (\hat{T}_h \hat{V}_{h+1})(s, a) + \Gamma_h(s, a) \geq 0.$$

**Step 2:** next we show $\zeta_h(s, a) \leq 2\Gamma_h(s, a), \forall s, a, h \in S \times A \times [H].$

Indeed, we have $\hat{Q}_h(s, a) = \max(Q_h(s, a), 0)$ and this is because: $Q_h(x, a) = (\hat{T}_h \hat{V}_{h+1})(x, a) - \Gamma_h(x, a) \leq (T_h \hat{V}_{h+1})(x, a) \leq H - h + 1$. Therefore, in this case we have:

$$\zeta_h(s, a) := (T_h \hat{V}_{h+1})(s, a) - \hat{Q}_h(s, a) \leq (T_h \hat{V}_{h+1})(s, a) - \hat{Q}_h(s, a)$$

$$= (T_h \hat{V}_{h+1})(s, a) - (\hat{T}_h \hat{V}_{h+1})(s, a) + \Gamma_h(s, a) \leq 2 \cdot \Gamma_h(s, a).$$

For the last statement, denote $\mathcal{F} := \{0 \leq \zeta_h(s, a) \leq 2\Gamma_h(s, a), \forall s, a, h \in S \times A \times [H]\}$. Note conditional on $\mathcal{F}$, then by equation 8, $V^\pi_1(s) - V^\pi_1(s) \leq \sum_{h=1}^{H} 2 \cdot E_\pi [\Gamma_h(s_h, a_h) \mid s_1 = s]$ holds for any policy $\pi$ almost surely. Therefore,

$$P \left[ \forall \pi, \ V^\pi_1(s) - V^\pi_1(s) \leq \sum_{h=1}^{H} 2 \cdot E_\pi [\Gamma_h(s_h, a_h) \mid s_1 = s] \right]$$

$$\geq P \left[ \forall \pi, \ V^\pi_1(s) - V^\pi_1(s) \leq \sum_{h=1}^{H} 2 \cdot E_\pi [\Gamma_h(s_h, a_h) \mid s_1 = s] \mid \mathcal{F} \right] \cdot P(\mathcal{F})$$

which finishes the proof.

C.2  BOUNDING $|(T_h \hat{V}_{h+1})(s, a) - (\hat{T}_h \hat{V}_{h+1})(s, a)|$.

By Lemma C.1, it remains to bound $|(T_h \hat{V}_{h+1})(s, a) - (\hat{T}_h \hat{V}_{h+1})(s, a)|$. Suppose $w_h$ is the coefficient corresponding to the $T_h \hat{V}_{h+1}$ (such $w_h$ exists by Lemma H.9), i.e. $T_h \hat{V}_{h+1} = \phi^T w_h$, and recall
\((\hat{T}_h \hat{V}_{h+1})(s, a) = \phi(s, a)^\top \hat{w}_h\), then:

\[
(\hat{T}_h \hat{V}_{h+1})(s, a) - (\hat{T}_h \hat{V}_{h+1})(s, a) = \phi(s, a)^\top (w_h - \hat{w}_h)
\]

\[
= \phi(s, a)^\top w_h - \phi(s, a)^\top \hat{A}_h^{-1} \left( \sum_{r=1}^{K} \phi(s^r_h, a^r_h) \cdot (r_h^r + \hat{V}_{h+1}(s^r_{h+1})) / \hat{\sigma}_h^2(s^r_h, a^r_h) \right)
\]

\[
= \phi(s, a)^\top w_h - \phi(s, a)^\top \hat{A}_h^{-1} \left( \sum_{r=1}^{K} \phi(s^r_h, a^r_h) \cdot (\hat{T}_h \hat{V}_{h+1})(s^r_h, a^r_h) / \hat{\sigma}_h^2(s^r_h, a^r_h) \right)
\]

\[
+ \phi(s, a)^\top \hat{A}_h^{-1} \left( \sum_{r=1}^{K} \phi(s^r_h, a^r_h) \cdot (r_h^r + \hat{V}_{h+1}(s^r_{h+1}) - (\hat{T}_h \hat{V}_{h+1})(s^r_h, a^r_h)) / \hat{\sigma}_h^2(s^r_h, a^r_h) \right).
\]

The term (i) is dealt by the following lemma.

**Lemma C.2.** Recall \(\kappa\) in Assumption 2.2. Suppose \(K \geq \max \{512H^4/\kappa^2 \log \left( \frac{2d}{\kappa} \right), 4\lambda H^2/\kappa \}\), then with probability \(1 - \delta\), for all \(s, a, h \in S \times A \times [H]\),

\[
\left| \phi(s, a)^\top w_h - \phi(s, a)^\top \hat{A}_h^{-1} \left( \sum_{r=1}^{K} \phi(s^r_h, a^r_h) \cdot (\hat{T}_h \hat{V}_{h+1})(s^r_h, a^r_h) / \hat{\sigma}_h^2(s^r_h, a^r_h) \right) \right| \leq \frac{2\lambda H^3 \sqrt{d}/\kappa}{K}.
\]

**Proof.** Recall \(T_h \hat{V}_{h+1} = \phi^\top w_h\) and apply Lemma H.6, we obtain with probability \(1 - \delta\), for all \(s, a, h \in S \times A \times [H]\),

\[
\phi(s, a)^\top w_h - \phi(s, a)^\top \hat{A}_h^{-1} \left( \sum_{r=1}^{K} \phi(s^r_h, a^r_h) \cdot (\hat{T}_h \hat{V}_{h+1})(s^r_h, a^r_h) / \hat{\sigma}_h^2(s^r_h, a^r_h) \right)
\]

\[
= \phi(s, a)^\top w_h - \phi(s, a)^\top \hat{A}_h^{-1} \left( \sum_{r=1}^{K} \phi(s^r_h, a^r_h) \cdot \phi(s^r_h, a^r_h)^\top w_h / \hat{\sigma}_h^2(s^r_h, a^r_h) \right)
\]

\[
= \phi(s, a)^\top w_h - \phi(s, a)^\top \hat{A}_h^{-1} \left( \hat{A}_h - \lambda I \right) w_h = \lambda \cdot \phi(s, a)^\top \hat{A}_h^{-1} w_h
\]

\[
\leq \lambda \|\phi(s, a)\|_{\hat{A}_h^{-1}} \cdot \|w_h\|_{\hat{A}_h^{-1}} \leq \frac{\lambda}{K} \|\phi(s, a)\|_{(\hat{A}_h^p)^{-1}} \cdot \|w_h\|_{(\hat{A}_h^p)^{-1}}
\]

\[
\leq \frac{\lambda}{K} \cdot \sqrt{\left\| (\hat{A}_h^p)^{-1} \right\| \cdot 2H^2 \sqrt{d} \cdot \sqrt{\left\| (\hat{A}_h^p)^{-1} \right\|}}
\]

where \(\hat{A}_h^p := \mathbb{E}_{h \sim P[h]} [\hat{\sigma}_h(s, a)^{-2} \phi(s, a) \phi(s, a)^\top]\) and the second inequality is by Lemma H.6 (with \(\phi' = \phi / \hat{\sigma}_h\) and \(\|\phi / \hat{\sigma}_h\| \leq \|\phi\| \leq 1 : = C\) and the third inequality uses \(\|a\| \cdot A \cdot a \leq \sqrt{\|a\|_2 \cdot \|A\|_2 \cdot \|a\|_2} = \|a\|_2 \sqrt{\|A\|_2}\) with \(a\) to be either \(\phi\) or \(w_h\). Moreover, \(\lambda_{\min}(\hat{A}_h^p) \geq \kappa / \max_{s, a} \hat{\sigma}_h(s, a)^{2} \geq \kappa / H^2\) implies \(\|\hat{A}_h^p\| \leq H^2 / \kappa\), therefore for all \(s, a, h \in S \times A \times [H]\), with probability \(1 - \delta\)

\[
\left| \phi(s, a)^\top w_h - \phi(s, a)^\top \hat{A}_h^{-1} \left( \sum_{r=1}^{K} \phi(s^r_h, a^r_h) \cdot (\hat{T}_h \hat{V}_{h+1})(s^r_h, a^r_h) / \hat{\sigma}_h^2(s^r_h, a^r_h) \right) \right| \leq \frac{2\lambda H^3 \sqrt{d}/\kappa}{K}.
\]

\(\square\)
For term (ii), denote: \( x_\tau = \frac{\phi(s_\tau, a_\tau)}{\hat{\sigma}(s_\tau, a_\tau)}. \) \( \eta_\tau = r_\tau + \hat{V}_{h+1}(s_{h+1}) - \left( T_h \hat{V}_{h+1}(s_h, a_h) / \hat{\sigma}(s_h, a_h) \right) \), then by Cauchy inequality it follows
\[
\left| \phi(s, a)^T \hat{\Lambda}_h^{-1} \left( \sum_{\tau=1}^{K} \phi(s_\tau, a_\tau) \cdot (r_\tau + \hat{V}_{h+1}(s_{h+1}) - \left( T_h \hat{V}_{h+1}(s_h, a_h) / \hat{\sigma}(s_h, a_h) \right) \right) \right| \\
\leq \sqrt{\phi(s, a)^T \hat{\Lambda}_h^{-1} \phi(s, a)} \cdot \| \sum_{\tau=1}^{K} x_\tau \eta_\tau \| \hat{\Lambda}_h^{-1} \tag{10}
\]

C.2.1 ANALYZING THE TERM \( \sqrt{\phi(s, a)\hat{\Lambda}_h^{-1} \phi(s, a)} \)

Recall (in Theorem 3.2) the estimated \( \hat{\Lambda}_h = \sum_{\tau=1}^{K} \phi(s_\tau, a_\tau) \phi(s_\tau, a_\tau)^T / \hat{\sigma}^2(s_\tau, a_\tau) + \lambda \cdot I \) and \( \Lambda_h = \sum_{\tau=1}^{K} \phi(s_\tau, a_\tau)^T \phi(s_\tau, a_\tau) / \sigma^2_{V_{h+1}}(s_\tau, a_\tau) + \lambda I \). Then we have the following lemma to control the term \( \sqrt{\phi(s, a)\hat{\Lambda}_h^{-1} \phi(s, a)} \).

**Lemma C.3.** Denote the quantities \( C_1 = \max\{2\lambda, 128 \log(2d/\delta), 128 H^4 \log(2d/\delta)/\kappa^2\} \) and \( C_2 = \max\{\frac{\lambda^2}{\log(1+K\lambda/2\delta)}, 96^2 H^{12} d \log((\lambda + K)/H/\delta)/\kappa^3\} \). Suppose the number of episode \( K \) satisfies \( K > \max\{C_1, C_2\} \), then with probability \( 1 - \delta \),
\[
\sqrt{\phi(s, a)\hat{\Lambda}_h^{-1} \phi(s, a)} \leq 2 \sqrt{\phi(s, a)\Lambda_h^{-1} \phi(s, a)} \quad \forall s, a \in S \times A.
\]

**Proof of Lemma C.3.** By definition \( \sqrt{\phi(s, a)\hat{\Lambda}_h^{-1} \phi(s, a)} = \| \phi(s, a) \| \hat{\Lambda}_h^{-1} \). Then denote
\[
\hat{\Lambda}_h = \frac{1}{K} \hat{\Lambda}_h, \quad \Lambda_h = \frac{1}{K} \Lambda_h,
\]
where \( \Lambda_h = \sum_{\tau=1}^{K} \phi(s_\tau, a_\tau)^T \phi(s_\tau, a_\tau) / \sigma^2_{V_{h+1}}(s_\tau, a_\tau) + \lambda I \). Under the condition of \( K \), by Lemma C.7, with probability \( 1 - \delta \)
\[
\| \hat{\Lambda}_h - \Lambda_h \| \leq \sup_{s,a} \left| \phi(s, a) \phi(s, a)^T / \sigma^2_{V_{h+1}}(s, a) - \phi(s, a) \phi(s, a)^T / \sigma^2_{V_{h+1}}(s, a) \right| \\
\leq \sup_{s,a} \left| \frac{\sigma^2_{V_{h+1}}(s, a)}{\sigma^2_{h}(s, a) \sigma^2_{V_{h+1}}(s, a)} \right| \cdot \| \phi(s, a) \|^2 \leq \sup_{s,a} \left| \frac{\sigma^2_{h}(s, a) - \sigma^2_{V_{h+1}}(s, a)}{\sigma^2_{V_{h+1}}(s, a)} \right| \cdot 1 \tag{11}
\]
\[
\leq 12 \sqrt{H^4 d \log \left( \frac{(\lambda + K)H}{\lambda \delta} \right)} + 12 \lambda H^2 \sqrt{d} / \kappa K.
\]

Next by Lemma H.5 (with \( \phi \) to be \( \phi / \sigma_{V_{h+1}} \) and \( C = 1 \)), it holds with probability \( 1 - \delta \),
\[
\| \Lambda_h - \left( \mathbb{E}_{\mu,h}[\phi(s, a) \phi(s, a)^T / \sigma^2_{V_{h+1}}(s, a)] + \frac{\lambda}{K} I_d \right) \| \leq \frac{4 \sqrt{2}}{\sqrt{K}} \left( \log \frac{2d}{\delta} \right)^{1/2}.
\]
Therefore by Weyl’s spectrum theorem and the condition $K > \max\{2\lambda, 128\log(2d/\delta), 128H^4\log(2d/\delta)/\kappa^2\}$, the above implies

$$
\|\Lambda_h^\prime\| = \lambda_{\max}(\Lambda_h^\prime) \leq \lambda_{\max}\left(\mathbb{E}_{\mu_h}[\phi(s, a)\phi(s, a)^\top/\sigma_{\mathcal{V}_{h+1}}^2(s, a)]\right) + \frac{\lambda}{K} + \frac{4\sqrt{2}}{\sqrt{K}}\left(\log\frac{2d}{\delta}\right)^{1/2}
$$

$$
= \left\|\mathbb{E}_{\mu_h}[\phi(s, a)\phi(s, a)^\top/\sigma_{\mathcal{V}_{h+1}}^2(s, a)]\right\|_2 + \frac{\lambda}{K} + \frac{4\sqrt{2}}{\sqrt{K}}\left(\log\frac{2d}{\delta}\right)^{1/2}
$$

$$
\leq \|\phi(s, a)\|^2 + \frac{\lambda}{K} + \frac{4\sqrt{2}}{\sqrt{K}}\left(\log\frac{2d}{\delta}\right)^{1/2} \leq 1 + \frac{\lambda}{K} + \frac{4\sqrt{2}}{\sqrt{K}}\left(\log\frac{2d}{\delta}\right)^{1/2} \leq 2,
$$

$$
\lambda_{\min}(\Lambda_h^\prime) \geq \lambda_{\min}\left(\mathbb{E}_{\mu_h}[\phi(s, a)\phi(s, a)^\top/\sigma_{\mathcal{V}_{h+1}}^2(s, a)]\right) + \frac{\lambda}{K} - \frac{4\sqrt{2}}{\sqrt{K}}\left(\log\frac{2d}{\delta}\right)^{1/2}
$$

$$
\geq \lambda_{\min}\left(\mathbb{E}_{\mu_h}[\phi(s, a)\phi(s, a)^\top/\sigma_{\mathcal{V}_{h+1}}^2(s, a)]\right) - \frac{4\sqrt{2}}{\sqrt{K}}\left(\log\frac{2d}{\delta}\right)^{1/2}
$$

$$
\geq \frac{\kappa}{H^2} - \frac{4\sqrt{2}}{\sqrt{K}}\left(\log\frac{2d}{\delta}\right)^{1/2} \geq \frac{\kappa}{2H^2}.
$$

Hence with probability $1 - \delta$, $\|\Lambda_h^\prime\| \leq 2$ and $\|\Lambda_h^{-1}\| = 1/\lambda_{\min}(\Lambda_h^\prime) \leq 2H^2/\kappa$. Similarly, one can show $\|\hat{\Lambda}_h^{-1}\| \leq 2H^2/\kappa$ with high probability.

Now apply Lemma H.4 to $\hat{\Lambda}_h^\prime$ and $\Lambda_h^\prime$ and a union bound, we obtain with probability $1 - \delta$, for all $s, a$

$$
\|\phi(s, a)\|_{\hat{\Lambda}_h^{-1}} \leq \left[1 + \sqrt{\|\Lambda_h^{-1}\| \|\hat{\Lambda}_h^{-1}\| \|\hat{\Lambda}_h - \Lambda_h^\prime\|} \cdot \|\phi(s, a)\|_{\Lambda_h^{-1}}
$$

$$
\leq \left[1 + \frac{2H^2}{\kappa} \cdot \frac{2H^2}{\kappa} \cdot \|\hat{\Lambda}_h - \Lambda_h^\prime\| \cdot \|\phi(s, a)\|_{\Lambda_h^{-1}}
$$

$$
\leq \left[1 + \frac{48H^4}{\kappa^2} \sqrt{H^4d \log\left(\frac{(\lambda + K)H}{\lambda\delta}\right)} + \frac{\lambda H^2\sqrt{d}}{\kappa K} \right] \cdot \|\phi(s, a)\|_{\Lambda_h^{-1}}
$$

$$
\leq \left[1 + \frac{96H^4}{\kappa^2} \sqrt{H^4d \log\left(\frac{(\lambda + K)H}{\lambda\delta}\right)} \right] \cdot \|\phi(s, a)\|_{\Lambda_h^{-1}} \leq 2 \|\phi(s, a)\|_{\Lambda_h^{-1}}
$$

where the third inequality uses equation 11 and the last and the second last inequality use $K > \max\{\kappa\log((\lambda + K)H/\lambda\delta), 96^2H^2d\log((\lambda + K)H/\lambda\delta)/\kappa^5\}$. Note the above is equivalent to $\sqrt{\phi(s, a)\hat{\Lambda}_h^{-1}\phi(s, a)} \leq 2\sqrt{\phi(s, a)\Lambda_h^{-1}\phi(s, a)}$ by multiplying $1/\sqrt{K}$ on both sides.

C.2.2 ANALYZING THE TERM $\|\sum_{r=1}^{K} x_r\eta_r\|_{\hat{\Lambda}_h}$

Lemma C.4. Recall $x_r = \frac{\phi(s_r, a_r)}{\sigma(s_r, a_r)}$ and $\eta_r = \left(r_h + \hat{\mathcal{V}}_{h+1} (s_r, a_r) - (\hat{T}_h \hat{\mathcal{V}}_{h+1}) (s_r, a_r)\right)/\tilde{\sigma}(s_r, a_r)$.

Let $C_{H,d,K} := 36\frac{H^4d}{\kappa} \log\left(\frac{(\lambda + K)2H^2d/\lambda\delta}{\lambda\delta}\right) + 12\lambda H^2\sqrt{d}$ and denote

$$
\xi := \sup_{V \in [0, H], s' \sim P_h(s,a), h \in [H]} \left| r_h + V(s') - (\hat{T}_h V) (s, a) \right| \sigma_V(s, a).
$$
If $K \geq 4C^2_{H,d,n,K}$ and $K \geq \tilde{O}(H^6d/K)$, then with probability $1 - \delta$,
\[
\left\| \sum_{\tau=1}^{K} x_{\tau} \eta_{\tau} \right\|_{\hat{\lambda}^{-1}} \leq 16 \sqrt{d \log \left( 1 + \frac{K}{\lambda d} \right) \cdot \log \left( \frac{4K^2}{\delta} \right) + 4\xi \log \left( \frac{4K^2}{\delta} \right)} \leq \tilde{O} \max \left\{ \sqrt{d}, \xi \right\},
\]
where $\tilde{O}$ absorbs the constants and Polylog terms.

**Proof of Lemma C.4.** By construction, we have $\|x\| \leq \|\hat{\sigma}\| \leq 1$ and by Lemma C.7, with probability $1 - \delta/3$,
\[
\left\| \sigma_{\hat{V}_{h+1}} - \hat{\sigma}_h \right\|_{\infty} = \sup_{s,a} \left\| \sigma_{\hat{V}_{h+1}}^2(s,a) - \hat{\sigma}_h^2(s,a) \right\| \leq \frac{1}{2} \left\| \sigma_{\hat{V}_{h+1}}^2(s,a) + \hat{\sigma}_h(s,a) \right\| \leq C_{H,d,n,K} \sqrt{\frac{1}{K}}
\]
Therefore, when $K \geq 4C^2_{H,d,n,K}, C_{H,d,n,K} \sqrt{\frac{1}{K}} \leq 1/2 \leq \sigma_{\hat{V}_{h+1}}(s_h^*, a_h^*)/2$ and hence
\[
|\eta| \leq \frac{r_h^* + \hat{V}_{h+1}(s_h^*) - (T_h \hat{V}_{h+1})(s_h^*, a_h^*)}{\sigma_{\hat{V}_{h+1}}^2(s_h^*, a_h^*) - C_{H,d,n,K} \sqrt{\frac{1}{K}} + 1} \leq 2 \left\| \frac{r_h^* + \hat{V}_{h+1}(s_h^*) - (T_h \hat{V}_{h+1})(s_h^*, a_h^*)}{\sigma_{\hat{V}_{h+1}}^2(s_h^*, a_h^*)} \right\| := \xi.
\]

Next, for a fixed function $V$, we define the Bellman error as $\mathcal{B}_h(V)(s, a) = r_h + V(s') - (T_h V)(s, a)$, then
\[
\text{Var}[\eta, \mathcal{F}_{\tau-1}] = \text{Var} \left[ \frac{r_h^* + \hat{V}_{h+1}(s_h^*) - (T_h \hat{V}_{h+1})(s_h^*, a_h^*)}{\sigma_{\hat{V}_{h+1}}^2(s_h^*, a_h^*)} \right| \mathcal{F}_{\tau-1} \right] = \frac{\text{Var} \left[ \mathcal{B}_h \hat{V}_{h+1}(s_h^*, a_h^*) - \mathcal{B}_h V_{h+1}(s_h^*, a_h^*) + \mathcal{B}_h V_{h+1}^*(s_h^*, a_h^*) \right]}{\sigma_{\hat{V}_{h+1}}^2(s_h^*, a_h^*)} \leq \frac{\text{Var} \left[ \mathcal{B}_h V_{h+1}^*(s_h^*, a_h^*) \right]}{\sigma_{\hat{V}_{h+1}}^2(s_h^*, a_h^*)} \leq \frac{\text{Var} \left[ \mathcal{B}_h V_{h+1}^*(s_h^*, a_h^*) \right]}{\sigma_{\hat{V}_{h+1}}^2(s_h^*, a_h^*)} + \tilde{O} \left( \frac{H^2 \sqrt{d}}{\sqrt{K}} \right) \leq \frac{\text{Var} \left[ \mathcal{B}_h V_{h+1}^*(s_h^*, a_h^*) \right]}{\sigma_{\hat{V}_{h+1}}^2(s_h^*, a_h^*)} + \tilde{O} \left( \frac{H^2 \sqrt{d}}{\sqrt{K}} \right) \leq \frac{2 \text{Var} \left[ V_{h+1}^*(s_h^*, a_h^*) \right]}{\sigma^2(s_h^*, a_h^*)} + \tilde{O} \left( \frac{H^2 \sqrt{d}}{\sqrt{K}} \right) \leq 2 \text{Var} \left[ V_{h+1}^*(s_h^*, a_h^*) \right] + \tilde{O} \left( \frac{H^2 \sqrt{d}}{\sqrt{K}} \right) \leq \tilde{O}(1)
\]
where the first inequality is by Lemma H.11, the second inequality is by $T_h$ is non-expansive, the third inequality is by Lemma C.8, the next equality is by Markovian property, and the fourth inequality is by Lemma C.7 and Lemma C.10. The fifth inequality uses definition $\sigma_{S_h,V}(s,a)^2 := \max\{1, \text{Var}_{S_h,V}(V)(s,a)\}$ and the last one is by condition $K \geq \tilde{O}(H^6d/K)$ and $\sigma_{S_h,V}(s,a)^2 := \max\{1, \text{Var}_{S_h,V}(V^*)(s,a)\} \geq 1$. Thus, by Bernstein inequality for self-normalized martingale.
Algorithm 1 satisfies 0

\[\sum_{\tau=1}^{K} x_\tau \eta_\tau \leq \tilde{O}\left(\sqrt{d \log \left(1 + \frac{K}{\lambda \delta}\right) \cdot \log \left(\frac{4K^2}{\delta}\right)}\right) + 4\xi \log \left(\frac{4K^2}{\delta}\right) \leq \tilde{O}\max\{\sqrt{d}, \xi\}\]

where \(\tilde{O}\) absorbs the constants and Polylog terms.

Recall \(M_1, M_2, M_3, M_4\) in List A. Based on the above results, we have the following key lemma:

**Lemma C.5.** Assume \(K > \max\{M_1, M_2, M_3, M_4\}\), for any \(0 < \lambda < \kappa\), suppose \(\sqrt{d} > \xi\), and \(\xi := \sup_{V \in [0, H], s' \sim P_h(s, a), h \in [H]} \left|\frac{r_h + V(s') - (T_h V)(s, a)}{\sigma_V(s, a)}\right|\). Then with probability \(1 - \delta\), for all \(h, s, a \in [H] \times S \times A\),

\[\left|(\T_h \V_{h+1} - \T_h \V_{h+1})(s, a)\right| \leq \tilde{O}\left(\sqrt{d} \sqrt{\phi(s, a)\Lambda_h^{-1}\phi(s, a)} + \frac{2H^3 \sqrt{d}}{K}\right),\]

where \(\Lambda_h = \sum_{\tau=1}^{K} \phi(s_h, a_h)^\top \phi(s_h, a_h) / \sigma_{\V_{h+1}}^2(s_h, a_h) + \lambda I\) and \(\tilde{O}\) absorbs the universal constants and Polylog terms.

**Proof of Lemma C.5.** Combing equation 9, Lemma C.2, equation 10, Lemma C.3 and C.4 and a union bound to finish the proof.

**C.3 Proof of the first part of Theorem 3.2**

**Theorem C.6** (First part of Theorem 3.2). Let \(K\) be the number of episodes. Suppose \(\sqrt{d} > \xi\), where \(\xi := \sup_{V \in [0, H], s' \sim P_h(s, a), h \in [H]} \left|\frac{r_h + V(s') - (T_h V)(s, a)}{\sigma_V(s, a)}\right|\) and \(K > \max\{M_1, M_2, M_3, M_4\}\). Then for any \(0 < \lambda < \kappa\), with probability \(1 - \delta\), for all policy \(\pi\) simultaneously, the output \(\hat{\pi}\) of Algorithm 1 satisfies

\[v^\pi - v^\\hat{\pi} \leq \tilde{O}\left(\sqrt{d} \cdot \sum_{h=1}^{H} E_{\pi} \left[(\phi(\cdot, \cdot)^\top \Lambda_h^{-1} \phi(\cdot, \cdot))^{1/2}\right]\right) + \frac{2H^4 \sqrt{d}}{K},\]

where \(\Lambda_h = \sum_{\tau=1}^{K} \frac{\phi(s_h, a_h)^\top \phi(s_h, a_h)}{\sigma_{\V_{h+1}}^2(s_h, a_h)} + \lambda I\) and \(\tilde{O}\) absorbs the universal constants and the Polylog terms.

**Proof of Theorem C.6.** Combing Lemma C.1 and Lemma C.5, we directly have with probability \(1 - \delta\), for all policy \(\pi\) simultaneously,

\[V^\pi(s) - V^\\hat{\pi}(s) \leq \tilde{O}\left(\sqrt{d} \cdot \sum_{h=1}^{H} E_{\pi} \left[(\phi(\cdot, \cdot)^\top \Lambda_h^{-1} \phi(\cdot, \cdot))^{1/2}\right]s_1 = s\right) + \frac{2H^4 \sqrt{d}}{K},\]

now take the initial distribution \(d_1\) on both sides to get the stated result.

**C.4 Two Intermediate Results**

The next two lemmas provide intermediate results in finishing the whole proofs.

---

\(^9\)To be rigorous, Lemma H.3 needs to be modified since the absolute value bound and the variance bound here are in the high probability sense. However, this will not affect the validity of the result as the weaker version can also be obtained (see Chung and Lu (2006) and a related discussion in Yin et al. (2021) Remark E.7.) To make the proof more readable, we do not include them here to avoid over-technicality.

\(^{10}\)The definition of \(M_1\) is in List A.
C.4.1 BOUNDING THE VARIANCE

Lemma C.7. Recall the definition \( \hat{\sigma}_h(\cdot, \cdot)^2 = \max\{1, \text{Var}_P \hat{h} \} \) and \( \sigma_{\hat{h}}(\cdot, \cdot)^2 := \max\{1, \text{Var}_P \hat{h} \} \). Moreover, \( \text{Var}_h \hat{h}^2 \) is \( \mathbb{E} \hat{h} \) where \( \hat{h} \) and \( \hat{h}^2 \) are defined in Algorithm 1. Let \( K \geq \max \{512(1/\kappa)^2 \log \left( \frac{Hd}{\delta} \right), 4\lambda/\kappa \} \), then with probability \( 1 - \delta \),
\[
\sup_h |\hat{\sigma}_h^2 - \sigma_{\hat{h}}^2| \leq 36 \sqrt{\frac{H^4d^3}{\kappa K} \log \left( \frac{(\lambda + K)2KdH^2}{\lambda \delta} \right) + 12\lambda H^2 \sqrt{d}}.
\]

Proof. Step 1: we first show for all \( h, s, a \in [H] \times S \times \mathcal{A} \), with probability \( 1 - \delta \)
\[
|\langle \phi(s, a), \beta_h \rangle_{[0, (H-h^2)]} - \mathbb{P}_h(\hat{h})^2(s, a)| \leq 12 \sqrt{\frac{H^4d^3}{\kappa K} \log \left( \frac{(\lambda + K)2KdH^2}{\lambda \delta} \right) + 4\lambda H^2 \sqrt{d}}.
\]

Proof of Step 1. Note
\[
\begin{align*}
|\langle \phi(s, a), \beta_h \rangle_{[0, (H-h^2)]} - \mathbb{P}_h(\hat{h})^2(s, a)| & = |\langle \phi(s, a), \beta_h \rangle - \mathbb{P}_h(\hat{h})^2(s, a)| \\
& = |\phi(s, a)^T \Sigma_{h}^{-1} \sum_{k=1}^{K} \phi(s_k, a_k) \cdot \hat{h}_{k+1}(s_k) - \mathbb{P}_h(\hat{h})^2(s, a)| \\
& = |\phi(s, a)^T \Sigma_{h}^{-1} \sum_{k=1}^{K} \phi(s_k, a_k) \cdot \hat{h}_{k+1}(s_k) - \phi(s, a)^T \Sigma_{h}^{-1} \sum_{k=1}^{K} \phi(s_k, a_k) \cdot \hat{h}_{k+1}(s_k)| \\
& = \lambda \left| \phi(s, a)^T \Sigma_{h}^{-1} \int_S (\hat{h})^2(s')d\nu_h(s') \right| + \lambda \left| \phi(s, a)^T \Sigma_{h}^{-1} \int_S (\hat{h})^2(s')d\nu_h(s') \right|.
\end{align*}
\]

For 2, since \( K \geq \max \{512(1/\kappa)^2 \log \left( \frac{Hd}{\delta} \right), 4\lambda/\kappa \} \), by Lemma H.6 and a union bound over \( h \in [H] \), with probability \( 1 - \delta \) for all \( h, s, a \in [H] \times S \times \mathcal{A} \),
\[
\begin{align*}
\mathbb{P}_h(\hat{h})^2(s, a) & \leq \lambda \left| \phi(s, a)^T \Sigma_{h}^{-1} \int_S (\hat{h})^2(s')d\nu_h(s') \right| + \lambda \left| \phi(s, a)^T \Sigma_{h}^{-1} \int_S (\hat{h})^2(s')d\nu_h(s') \right| \\
& \leq \frac{2}{\sqrt{K}} \left| \phi(s, a)^T \Sigma_{h}^{-1} \int_S (\hat{h})^2(s')d\nu_h(s') \right| + \frac{2}{\sqrt{K}} \left| \phi(s, a)^T \Sigma_{h}^{-1} \int_S (\hat{h})^2(s')d\nu_h(s') \right| \\
& \leq 4\lambda \left( \Sigma_{h}^{-1} \right)^{-1} \left| \frac{H^2 \sqrt{d}}{K} \right| \leq 4\lambda \left( \frac{H^2 \sqrt{d}}{\kappa K} \right).
\end{align*}
\]

For 1, we have
\[
\begin{align*}
\mathbb{P}_h(\hat{h})^2(s, a) & \leq \lambda \left| \phi(s, a)^T \Sigma_{h}^{-1} \sum_{k=1}^{K} \phi(s_k, a_k) \cdot \hat{h}_{k+1}(s_k) - \mathbb{P}_h(\hat{h})^2(s, a) \right| \\
& \leq \frac{2}{\sqrt{K}} \left| \phi(s, a)^T \Sigma_{h}^{-1} \int_S (\hat{h})^2(s')d\nu_h(s') \right| + \frac{2}{\sqrt{K}} \left| \phi(s, a)^T \Sigma_{h}^{-1} \int_S (\hat{h})^2(s')d\nu_h(s') \right| \\
& \leq 8H^4 \cdot \frac{d}{\delta} \log \left( \frac{\lambda + K}{\lambda \delta} \right).
\end{align*}
\]

Bounding using covering. Note for any fix \( V_h+1 \), we can define \( x_r = \phi(s, a) \) and \( \eta_r = (\hat{h})^2(s, a) \) is \( H^2 \)-subgaussian, by Lemma H.2 (where \( t = K \) and \( L = 1 \)) with probability 1 - \( \delta \),
\[
\left| \sum_{r=1}^{K} \phi(s, a) \cdot (\hat{h})^2(s, a) \right| \leq \sqrt{sH^4 \cdot \frac{d}{\delta} \log \left( \frac{\lambda + K}{\lambda \delta} \right)}.
\]

let \( N_h(\epsilon) \) be the minimal \( \epsilon \)-cover (with respect the supremum norm) of \( V_h := \{ V_h : V_h(\cdot) = \max_{a \in \mathcal{A}} \left\{ \min\{\phi(s, a)^T \theta - C_1 \sqrt{d} \cdot \phi(s, a)^T \Lambda_h^{-1} \phi(s, a) - C_2, H - h + 1 \} \right\} \} \). That is, for any
\( V \in V_h \), there exists a value function \( V' \in \mathcal{N}_h(\epsilon) \) such that \( \sup_{s \in S} |V(s) - V'(s)| < \epsilon \). Now by a union bound, we obtain with probability \( 1 - \delta \)

\[
\sup_{V_{h+1} \in \mathcal{N}_{h+1}(\epsilon)} \left| \sum_{\tau=1}^{K} \phi(s_h^\tau, a_h^\tau) \cdot (V_{h+1}(s_{h+1}^\tau)^2 - \mathbb{P}_h(V_{h+1})^2(s_h^\tau, a_h^\tau)) \right|_{\Sigma_h^{-1}} \leq \sqrt{8H^4 \cdot \frac{d}{2} \log \left( \frac{\lambda + K}{\lambda \delta} |\mathcal{N}_{h+1}(\epsilon)| \right)}
\]

which implies

\[
\left| \sum_{\tau=1}^{K} \phi(s_h^\tau, a_h^\tau) \cdot \left( \hat{V}_{h+1}(s_{h+1}^\tau)^2 - \mathbb{P}_h(\hat{V}_{h+1})^2(s_h^\tau, a_h^\tau) \right) \right|_{\Sigma_h^{-1}} \leq \sqrt{8H^4 \cdot \frac{d}{2} \log \left( \frac{\lambda + K}{\lambda \delta} |\mathcal{N}_{h+1}(\epsilon)| \right)} + 4H^2 \sqrt{d^2/\lambda}
\]

choosing \( \epsilon = d\sqrt{\lambda}/K \), applying Lemma B.3 of Jin et al. (2021b)\(^1\) to the covering number \( \mathcal{N}_{h+1}(\epsilon) \) w.r.t. \( V_{h+1} \), we can further bound above by

\[
\leq \sqrt{8H^4 \cdot \frac{d^3}{2} \log \left( \frac{\lambda + K}{\lambda \delta} 2dHK \right)} + 4H^2 \sqrt{d^2} \leq 6H^4 \cdot \frac{d^3}{2} \log \left( \frac{\lambda + K}{\lambda \delta} 2dHK \right)
\]

Apply a union bound for \( h \in [H] \), we have with probability \( 1 - \delta \), for all \( h \in [H] \),

\[
\left| \sum_{\tau=1}^{K} \phi(s_h^\tau, a_h^\tau) \cdot \left( \hat{V}_{h+1}(s_{h+1}^\tau)^2 - \mathbb{P}_h(\hat{V}_{h+1})^2(s_h^\tau, a_h^\tau) \right) \right|_{\Sigma_h^{-1}} \leq 6 \sqrt{H^4 d^3 \log \left( \frac{(\lambda + K)2KdH^2}{\lambda \delta} \right)} \tag{15}
\]

and similar to (\( \mathcal{Q} \)), with probability \( 1 - \delta \) for all \( h, s, a \in [H] \times S \times A \),

\[
\| \phi(s, a) \|_{\Sigma_h^{-1}} \leq 2 \frac{\| (\Sigma^p)_h^{-1} \|_{\Sigma_h^{-1}}^{1/2}}{\sqrt{K}} \leq \frac{2}{\sqrt{\kappa K}} \tag{16}
\]

Combing equation 13, equation 14, equation 15 and equation 16 we obtain with probability \( 1 - \delta \) for all \( h, s, a \in [H] \times S \times A \),

\[
\left| \langle \phi(s, a), \hat{\beta}_h \rangle_{[0,(H-h+1)^2]} - \mathbb{P}_h(\hat{V}_{h+1})^2(s, a) \right| \leq 12 \frac{H^4 d^3}{\kappa K} \log \left( \frac{(\lambda + K)2KdH^2}{\lambda \delta} \right) + 4\lambda \frac{H^2 \sqrt{d}}{\kappa K} \tag{17}
\]

The proof of Step2 follows nearly the identical way as Step1 except \( \hat{V}_h^2 \) is replaced by \( \hat{V}_h \).

**Step3:** We prove \( \sup_h \| \hat{\sigma}_h^2 - \hat{\sigma}_h^2 \| \leq 36 \frac{H^4 d^3 \log \left( \frac{(\lambda + K)2KdH^2}{\lambda \delta} \right)}{\kappa K} + 12\lambda \frac{H^2 \sqrt{d}}{\kappa K} \).

**Proof of Step3.** By equation 17,

\[
\left| \langle \phi(s, a), \hat{\beta}_h \rangle_{[0,(H-h+1)^2]} - \mathbb{P}_h(\hat{V}_{h+1})^2(s, a) \right| \leq 12 \frac{H^4 d^3}{\kappa K} \log \left( \frac{(\lambda + K)2KdH^2}{\lambda \delta} \right) + 4\lambda \frac{H^2 \sqrt{d}}{\kappa K} \tag{17}
\]

\(^1\)Note the same result in Jin et al. (2021b) applies even though we have an extra constant \( C_2 \).
Combining this with Step1 we receive \( \forall h, s, a \in [H] \times S \times A \), with probability \( 1 - \delta \)
\[
\left| \widehat{\text{Var}}_h \hat{V}_{h+1}(s, a) - \text{Var}_h \hat{V}_{h+1}(s, a) \right| \leq 36 \sqrt{\frac{H^4 d^3}{\kappa K}} \log \left( \frac{\frac{1}{\lambda} + K \cdot d H^2}{\sqrt{\lambda \delta}} \right) + 12 \frac{H^2 \sqrt{d}}{\kappa K}.
\]
Finally, by the non-expansiveness of operator \( \max \{1, \cdot \} \), we have the stated result.

\[
\square
\]

C.4.2 A CRUDE BOUND ON \( \sup_h ||V^*_h - \hat{V}_h||_\infty \).

**Lemma C.8.** Define \( \overline{\sigma}_h(s, a) = \sqrt{\max \left\{ 1, \text{Var}_h \hat{V}_{h+1}(s, a) \right\} + 1} \), if \( K \geq \max \{M_1, M_2, M_3, M_4 \} \) and \( K > C \cdot H^4 \kappa^2 \), then with probability at least \( 1 - \delta \),
\[
\sup_h ||V^*_h - \hat{V}_h||_\infty \leq \tilde{O} \left( \frac{H^2 \sqrt{d}}{\sqrt{\kappa K}} \right).
\]

**Proof.** **Step1:** We show with probability at least \( 1 - \delta \), \( \sup_h \|V^*_h - V^\overline{\sigma}_h\|_\infty \leq \tilde{O} \left( \frac{H^2 \sqrt{d}}{\sqrt{\kappa K}} \right) \).

Next, since \( K \geq \max \{512(1/\kappa)^2 \log(\frac{4Hd}{\delta}), 4\lambda/\kappa \} \), by Lemma H.6 and a union bound over \( h \in [H] \), with probability \( 1 - \delta \)
\[
\sup_{s, a} \|\phi(s, a)\|_{\Lambda_h^{-1}} \leq \frac{2}{\sqrt{\kappa K}} \sup_{s, a} \|\phi(s, a)\|_{\Lambda_h^{-1}} \leq \frac{2H}{\sqrt{\kappa K}}, \ \forall h \in [H].
\]

Lastly, taking \( \pi = \pi^* \) in equation 18 to obtain
\[
0 \leq V^\pi_h(s) - V^\overline{\sigma}_h(s) \leq \tilde{O} \left( \sqrt{d} \cdot \sum_{t=h}^{H} \mathbb{E}_{\pi^*} \left[ (\phi(\cdot, \cdot)^T \Lambda_t^{-1} \phi(\cdot, \cdot))^{1/2} \big| s_h = s \right] \right) + \frac{2H^4 \sqrt{d}}{K},
\]
\[
\leq \tilde{O} \left( \frac{H^3 \sqrt{d}}{\sqrt{\kappa K}} \right) + \frac{2H^4 \sqrt{d}}{K}.
\]

This implies by using the condition \( K > C \cdot H^4 \kappa^2 \), we finish the proof of Step1.

**Step2:** We show with probability \( 1 - \delta \), \( \sup_h \|\hat{V}_h - V^\overline{\sigma}_h\|_\infty \leq \tilde{O} \left( \frac{H^2 \sqrt{d}}{\sqrt{\kappa K}} \right) \).

Indeed, applying Extended Value Difference Lemma H.7 for \( \pi = \pi' = \hat{\pi} \), then with probability \( 1 - \delta \), for all \( s, h \)
\[
\left| \hat{V}_h(s) - V^\overline{\sigma}_h(s) \right| = \sum_{t=h}^{H} \mathbb{E}_\pi \left[ \hat{Q}_h(s_h, a_h) - \left( \hat{T}_h \hat{V}_{h+1} \right)(s_h, a_h) \big| s_h = s \right] \]
\[
\leq \sum_{t=h}^{H} \left\| (\hat{T}_h \hat{V}_{h+1} - T_h \hat{V}_{h+1})(s, a) \right\| + ||\Gamma_h(s, a)||
\]
\[
\leq \tilde{O} \left( H\sqrt{d} \right)^{\frac{1}{2} \phi(s, a) \Lambda_h^{-1} \phi(s, a)} + \frac{4H^4 \sqrt{d}}{K} \leq \tilde{O} \left( \frac{H^2 \sqrt{d}}{\sqrt{\kappa K}} \right).
\]
where the second inequality uses Lemma C.5\textsuperscript{12} and the last inequality follows the same procedure as Step1.

**Step3:** Combine Step1 and Step2, by triangular inequality and a union bound we finish the proof of the lemma.

\[ \frac{H^2 \sqrt{d}}{\sqrt{\kappa K}} \]

\[ \frac{d}{\kappa} \]

\[ \frac{d/\kappa}{\sqrt{\kappa}} \]

\[ O \left( \frac{H^2 \sqrt{d}}{\sqrt{\kappa K}} \right) \]

\[ C.5 \]

\[ \lambda_h \]

\[ \phi(s, a) \Lambda_h^{-1} \phi(s, a) \]

\[ \| \phi(s, a) \|_{\Lambda_h^{-1}} \]

\[ \Lambda_h' = \frac{1}{K} \Lambda_h, \quad \Lambda_h'' = \frac{1}{K} \Lambda_h', \]

\[ \text{for all } h \in [H], \text{ where the last inequality comes from Lemma C.8. Combining this with Lemma C.7, we have the stated result.} \]

\[ \textbf{Remark C.9.} \text{ Note as an intermediate calculation, equation 19 ensures a learning bound with order } O \left( \frac{H^2 \sqrt{d}}{\sqrt{\kappa K}} \right). \text{ Here, the convergence rate is the standard statistical rate } \frac{1}{\sqrt{K}} \text{ and the } H^2 \text{ dependence is loose. However, the feature dependence } \sqrt{d/\kappa} \text{ is roughly tight, since, in the well-explored case (Assumption 2 of Wang et al. (2021a)), } \kappa = 1/d \text{ and the } \sqrt{d/\kappa} = \sqrt{d^2} \text{ recovers the optimal feature dependence } dH \sqrt{T} \text{ in the online setting (Zhou et al., 2021a). If } \kappa \ll 1/d, \text{ then doing offline learning requires sample size proportional to } d/\kappa, \text{ which reveals offline RL is harder when the exploration of behavior policy is insufficient. When } \kappa = 0, \text{ learning the optimal policy accurately cannot be guaranteed even if the sample/episode size } K \to \infty. \]

C.5  PROOF OF THE SECOND PART OF THEOREM 3.2

\[ \textbf{Lemma C.10.} \text{ Recall } \tilde{\sigma}_h = \sqrt{\max \{ 1, \text{Var}_h \hat{V}_{h+1} \}} \text{ and } \sigma^*_h = \sqrt{\max \{ 1, \text{Var}_h V^*_{h+1} \}} + 1. \text{ Let } K \geq \max \{ 512(1/\kappa)^2 \log \left( \frac{dH}{\sigma} \right), 4\lambda/\kappa \} \text{ and } K \geq \max \{ M_1, M_2, M_3, M_4 \}, \text{ then with probability } 1 - \delta, \]

\[ \sup_h \| \sigma^2_h - \sigma^*_h \|_{\infty} \leq \tilde{O} \left( \frac{H^3 \sqrt{d}}{\sqrt{\kappa K}} \right). \]

\[ \textbf{Proof.} \text{ By definition and the non-expansiveness of } \max \{ 1, \cdot \} + 1, \text{ we have } \\
\]

\[ \| \sigma^2_h - \sigma^*_h \|_{\infty} \leq \| \text{Var}_h \hat{V}_{h+1} - \text{Var}_h V^*_{h+1} \|_{\infty} \]

\[ \leq \| \text{Var}_h (\hat{V}_{h+1}^2 - V^*_{h+1}^2) \|_{\infty} + \| (\text{Var}_h \hat{V}_{h+1})^2 - (\text{Var}_h V^*_{h+1})^2 \|_{\infty} \]

\[ \leq \| \hat{V}_{h+1}^2 - V^*_{h+1}^2 \|_{\infty} + \| (\text{Var}_h \hat{V}_{h+1} + \text{Var}_h V^*_{h+1})(\text{Var}_h \hat{V}_{h+1} - \text{Var}_h V^*_{h+1}) \|_{\infty} \]

\[ \leq 2H \| \hat{V}_{h+1} - V^*_{h+1} \|_{\infty} + 2H \| \text{Var}_h \hat{V}_{h+1} - \text{Var}_h V^*_{h+1} \|_{\infty} \leq \tilde{O} \left( \frac{H^3 \sqrt{d}}{\sqrt{\kappa K}} \right). \]

with probability \( 1 - \delta \) for all \( h \in [H] \), where the second inequality comes from Lemma C.8. Combining this with Lemma C.7, we have the stated result. \( \Box \)

\[ \textbf{Lemma C.11.} \text{ Denote the quantities } C_1 = \max \{ 2\lambda, 128 \log(2d/\delta), 128H^4 \log(2d/\delta)/\kappa^2 \} \text{ and } C_2 = \max \{ \frac{1}{2 \log((\lambda + K)H/\delta)}, \frac{96}{2} H^2 d \log((\lambda + K)H/\delta)/\kappa^2 \}. \text{ Suppose the number of episode } K \text{ satisfies } K > \max \{ C_1, C_2 \}, \text{ then with probability } 1 - \delta, \]

\[ \sqrt{\phi(s, a) \Lambda_h^{-1} \phi(s, a)} \leq 2 \sqrt{\phi(s, a) \Lambda_h^{*1} \phi(s, a)}, \quad \forall s, a \in S \times A. \]

\[ \sqrt{\phi(s, a) \Lambda_h^{-1} \phi(s, a)} = \| \phi(s, a) \|_{\Lambda_h^{-1}}. \]

\[ \textbf{Proof of Lemma C.11.} \text{ By definition } \sqrt{\phi(s, a) \Lambda_h^{-1} \phi(s, a)} = \| \phi(s, a) \|_{\Lambda_h^{-1}}. \text{ Then denote } \\
\]

\[ \Lambda_h' = \frac{1}{K} \Lambda_h, \quad \Lambda_h'' = \frac{1}{K} \Lambda_h'. \]

\[ \text{To be absolutely rigorous, we cannot directly apply Lemma C.5 here since the crude bound has already been used in Lemma C.4. However, this can be resolved completely by first deriving an even cruder bound for } \sup_h \| \hat{V}_h^2 - \hat{V}_h \|_{\infty} \text{ that has } 1/\sqrt{K} \text{ rate without using Lemma C.5 (which we call it Lemma C.8'), and we can use Lemma C.8} \text{ to show a similar result Lemma C.5". Finally, we can use Lemma C.5" here to finish the proof of this Lemma C.8. However, we avoid explicitly doing this to prevent over-technicality.} \]
where $\Lambda_h = \sum_{t=1}^{K} \phi(s_h^t, a_h^t)^T \phi(s_h^t, a_h^t) / \sigma^2_{V_h+1}(s_h^t, a_h^t) + \lambda I$. Under the condition of $K$, by Lemma C.10, with probability $1 - \delta$,

$$
\|\Lambda_h^* - \Lambda_h^1\| \leq \sup_{s, a} \left\| \frac{\phi(s, a) \phi(s, a)^T}{\sigma^2_h(s, a)} - \frac{\phi(s, a) \phi(s, a)^T}{\sigma^2_{V_h+1}(s, a)} \right\| \\
\leq \sup_{s, a} \left\| \sigma^2_h(s, a) - \sigma^2_{V_h+1}(s, a) \right\| \cdot \|\phi(s, a)||^2 \leq \sup_{s, a} \left\| \sigma^2_h(s, a) - \sigma^2_{V_h+1}(s, a) \right\| \cdot 1 \quad (20)
$$

Next by Lemma H.5 (with $\phi$ to be $\phi/\sigma_{V_h+1}$ and $C = 1$), it holds with probability $1 - \delta$,

$$
\left\| \Lambda_h^1 - \left( \mathbb{E}_{\mu, h}[\phi(s, a) \phi(s, a)^T / \sigma^2_{V_h+1}(s, a)] + \frac{\lambda}{K} I \right) \right\| \leq \frac{4\sqrt{2}}{\sqrt{K}} \left( \log \left( \frac{2d}{\delta} \right) \right)^{1/2}.
$$

Therefore by Weyl’s spectrum theorem and the condition $K > \max\{2\lambda, 128 \log(2d/\delta), 128H^4 \log(2d/\delta)/\kappa^2\}$, the above implies

$$
\left\| \Lambda_h^* \right\| = \lambda_{\max}(\Lambda_h^*) \leq \lambda_{\max} \left( \mathbb{E}_{\mu, h}[\phi(s, a) \phi(s, a)^T / \sigma^2_{V_h+1}(s, a)] \right) + \frac{\lambda}{K} + \frac{4\sqrt{2}}{\sqrt{K}} \left( \log \left( \frac{2d}{\delta} \right) \right)^{1/2} \\
\leq \left\| \mathbb{E}_{\mu, h}[\phi(s, a) \phi(s, a)^T / \sigma^2_{V_h+1}(s, a)] \right\| + \frac{\lambda}{K} + \frac{4\sqrt{2}}{\sqrt{K}} \left( \log \left( \frac{2d}{\delta} \right) \right)^{1/2} \\
\leq \|\phi(s, a)||^2 + \frac{\lambda}{K} + \frac{4\sqrt{2}}{\sqrt{K}} \left( \log \left( \frac{2d}{\delta} \right) \right)^{1/2} \leq 1 + \frac{\lambda}{K} + \frac{4\sqrt{2}}{\sqrt{K}} \left( \log \left( \frac{2d}{\delta} \right) \right)^{1/2} \leq 2,
$$

$$
\lambda_{\min}(\Lambda_h^*) \geq \lambda_{\min} \left( \mathbb{E}_{\mu, h}[\phi(s, a) \phi(s, a)^T / \sigma^2_{V_h+1}(s, a)] \right) + \frac{\lambda}{K} - \frac{4\sqrt{2}}{\sqrt{K}} \left( \log \left( \frac{2d}{\delta} \right) \right)^{1/2} \\
\geq \lambda_{\min} \left( \mathbb{E}_{\mu, h}[\phi(s, a) \phi(s, a)^T / \sigma^2_{V_h+1}(s, a)] \right) - \frac{4\sqrt{2}}{\sqrt{K}} \left( \log \left( \frac{2d}{\delta} \right) \right)^{1/2} \\
\geq \frac{\kappa}{H^2} - \frac{4\sqrt{2}}{\sqrt{K}} \left( \log \left( \frac{2d}{\delta} \right) \right)^{1/2} \geq \frac{\kappa}{2H^2}.
$$

Hence with probability $1 - \delta$, $\|\Lambda_h^*\| \leq 2$ and $\|\Lambda_h^{*-1}\| = 1/\lambda_{\min}(\Lambda_h^*) \leq 2H^2/\kappa$. Similarly, $\|\Lambda_h^{*-1}\| \leq 2H^2/\kappa$ with high probability.

Now apply Lemma H.4 to $\Lambda_h^*$ and $\Lambda_h^1$ and a union bound, we obtain with probability $1 - \delta$, for all $s, a$

$$
\|\phi(s, a)||_{\Lambda_h^{*-1}} \leq 1 + \sqrt{\|\Lambda_h^{*-1}\| \cdot \|\Lambda_h^*\| \cdot \|\Lambda_h^{*-1}\|} \cdot \|\phi(s, a)||_{\Lambda_h^{*-1}} \\
\leq 1 + \frac{2H^2}{\kappa} \cdot 1 \cdot \frac{2H^2}{\kappa} \cdot \|\Lambda_h^* - \Lambda_h^1\| \cdot \|\phi(s, a)||_{\Lambda_h^{*-1}} \\
\leq 1 + \frac{H^4}{\kappa^2} \left[ \tilde{O} \left( \frac{H^3\sqrt{d}}{\sqrt{K}} \right) \right] \cdot \|\phi(s, a)||_{\Lambda_h^{*-1}} \leq 2 \|\phi(s, a)||_{\Lambda_h^{*-1}}
$$

where the third inequality uses equation 20 and the last inequality uses $K > \max\{\kappa \log((\lambda + K)H/\lambda \delta), 96^2H^4 \log((\lambda + K)H^2/\lambda \delta)/\kappa^5\}$. The claimed result follows straightforwardly by multiplying $1/\sqrt{K}$ on both sides of the above.
Proof of Theorem 3.2. The first part of the theorem has been shown in Theorem C.6. For the second part, apply Theorem C.6 with \( \pi = \pi^* \), then with probability \( 1 - \delta \),

\[
v^\pi^* - v^\pi \leq \tilde{O} \left( \sqrt{d} \cdot \sum_{h=1}^{H} \mathbb{E}_{\pi^*} \left[ (\phi(\cdot, \cdot)^T \Lambda^{-1}_h \phi(\cdot, \cdot))^{1/2} \right] \right) + \frac{2H^4 \sqrt{d}}{K}.
\]

Now apply Lemma C.11 and a union bound, with probability \( 1 - \delta \),

\[
0 \leq v^* - v^\pi \leq \tilde{O} \left( \sqrt{d} \cdot \sum_{h=1}^{H} \mathbb{E}_{\pi^*} \left[ (\phi(\cdot, \cdot)^T \Lambda^{-1}_h \phi(\cdot, \cdot))^{1/2} \right] \right) + \frac{2H^4 \sqrt{d}}{K}.
\]

\( \square \)

## D PROOF OF THEOREM 3.3

First of all, we show the following lemma.

**Lemma D.1.** Suppose \( K > \max\{M_1, M_2, M_3, M_4\} \). Plug

\[
\Gamma_h^I(s, a) \leftarrow \phi(s, a)^T \tilde{\Lambda}_h^{-1} \sum_{\tau=1}^{K} \phi(s_h^\tau, a_h^\tau) \cdot \left( r_h^\tau + \tilde{V}_{h+1}(s_{h+1}^\tau) - (\tilde{T}_h \tilde{V}_{h+1})(s_h^\tau, a_h^\tau) \right) / \tilde{\sigma}_h^2(s_h^\tau, a_h^\tau) + \tilde{O}(\frac{H^3 d/\kappa}{K})
\]

in Algorithm 1 and let \( \tilde{T}_h \) be the Bellman operator and \( \tilde{V}_h \) be the approximated Bellman operator. Then we have with probability \( 1 - \delta \):

\[
|\langle \tilde{T}_h \tilde{V}_{h+1} - \tilde{T}_h \tilde{V}_{h+1} \rangle(s, a) | \leq \Gamma_h^I(s, a), \quad \forall s, a \in S \times A.
\]

**Proof of Lemma D.1.** Suppose \( w_h \) is the coefficient corresponding to the \( \tilde{T}_h \tilde{V}_{h+1} \) (such \( w_h \) exists by Lemma H.9), i.e. \( \tilde{T}_h \tilde{V}_{h+1} = \phi^T w_h \), and recall \( \langle \tilde{T}_h \tilde{V}_{h+1} \rangle(s, a) = \phi(s, a)^T \tilde{w}_h \), then:

\[
\left( \tilde{T}_h \tilde{V}_{h+1} \right)(s, a) - \left( \tilde{T}_h \tilde{V}_{h+1} \right)(s, a) = \phi(s, a)^T (w_h - \tilde{w}_h)
\]

\[
= \phi(s, a)^T w_h - \phi(s, a)^T \tilde{w}_h
\]

\[
= \phi(s, a)^T w_h - \phi(s, a)^T \tilde{w}_h
\]

\[
= \phi(s, a)^T w_h - \phi(s, a)^T \tilde{w}_h
\]

\[
= \phi(s, a)^T w_h - \phi(s, a)^T \tilde{w}_h
\]

\[
= \left( \phi(s, a)^T \tilde{w}_h \right) - \left( \phi(s, a)^T w_h \right)
\]

\[
= \phi(s, a)^T \tilde{w}_h - \phi(s, a)^T w_h
\]

For term (i), by Lemma C.2 it is bounded by \( \frac{2H^3 \sqrt{d}}{K} \) with probability \( 1 - \delta/2 \). For term (ii), it is bounded by

\[
\phi(s, a)^T \tilde{w}_h - \phi(s, a)^T w_h \leq \frac{2H^3 \sqrt{d}}{K}.
\]

\( \text{Note Here Lemma C.2 still applies even if the } \Gamma_h \text{ changes since it works for all } \tilde{V}_h \in [0, H] \text{ so that } \| w_h \|_2 \leq 2H \sqrt{d} \text{ and the truncation (Line 13 in Algorithm 1)} \text{ guarantees this.} \)

28
For term (iii), by Cauchy inequality
\[\phi(s,a) = \sum_{i=1}^{K} \phi(s_{h,i}, a_{h,i}) \cdot \left( (\tilde{T}_h \tilde{V}_{h+1}) (s_{h,i}, a_{h,i}) - (\tilde{T}_h \tilde{V}_{h+1}) (s_{h,i}, a_{h,i}) / \hat{\sigma}_h^2(s_{h,i}, a_{h,i}) \right)\]
\[\leq \|\phi(s,a)\|_{\hat{\Lambda}_h^{-1}} \cdot \sum_{i=1}^{K} \phi(s_{h,i}, a_{h,i}) \cdot \left( (\tilde{T}_h \tilde{V}_{h+1}) (s_{h,i}, a_{h,i}) - (\tilde{T}_h \tilde{V}_{h+1}) (s_{h,i}, a_{h,i}) / \hat{\sigma}_h^2(s_{h,i}, a_{h,i}) \right) \]
\[\leq \frac{2H}{\sqrt{\kappa \kappa}} \cdot \sqrt{d} \cdot \sqrt{d} \leq \hat{O}\left(\frac{H^2 d}{\kappa K}\right)
\]
where the first inequality is by Lemma H.6 (with \(\phi' = \phi / \tilde{\sigma}_h\) and \(\|\phi / \tilde{\sigma}_h\| \leq \|\phi\| \leq 1 := C\)) and the third inequality uses \(\sqrt{a^1} \cdot A \cdot \bar{a} \leq \|a_1\|_2 \|A\|_2 \|a\|_2 = \|a_1\|_2 \sqrt{\|A\|_2}\) with \(a\) to be either \(\phi\) or \(w_h\).
Moreover, \(\lambda_{\min}(\hat{\Lambda}_h) \geq \kappa / \max_{h,s,a} \tilde{\sigma}_h(s,a)^2 \geq \kappa / H^2\) implies \(\|\hat{\Delta}_h^{-1}\| \leq H^2 / \kappa\).

The second inequality is true by denoting \(x = \phi(s_{h,i}, a_{h,i}) / \hat{\sigma}(s_{h,i}, a_{h,i})\) and
\[\eta = \left( (\tilde{T}_h \tilde{V}_{h+1}) (s_{h,i}, a_{h,i}) - (\tilde{T}_h \tilde{V}_{h+1}) (s_{h,i}, a_{h,i}) / \hat{\sigma}_h(s_{h,i}, a_{h,i}) \right)
and use Lemma H.10 as the condition for applying Lemma H.2. By collecting those three terms together we have the result. \(\square\)

### D.1 Proof of Theorem 3.3

**Proof.** Use Lemma D.1 as the condition for Lemma C.1 and average over initial distribution \(d_1\), we obtain with probability \(1 - \delta\),
\[v^\pi - v^\tilde{\pi} \leq \sum_{h=1}^{H} \mathbb{E}_{\pi_h} [\phi(s,a)]^T \hat{\Lambda}_h^{-1} \sum_{r=1}^{K} \phi(s_{h,r}, a_{h,r}) \cdot \left( [T_h \tilde{V}_{h+1}] (s_{h,r}, a_{h,r}) - (T_h \tilde{V}_{h+1}) (s_{h,r}, a_{h,r}) / \hat{\sigma}_h^2(s_{h,r}, a_{h,r}) \right)
\]
\[(\phi) \leq \frac{O(H^2 d / \kappa)}{\kappa K}
\]

Denote \(A_h := \sum_{r=1}^{K} \phi(s_{h,r}, a_{h,r}) / \hat{\sigma}_h^2(s_{h,r}, a_{h,r})\), then
\[\mathbb{E}_{\pi_h} [\phi(s,a)]^T \hat{\Lambda}_h^{-1} \sum_{r=1}^{K} \phi(s_{h,r}, a_{h,r}) \cdot \left( [T_h \tilde{V}_{h+1}] (s_{h,r}, a_{h,r}) - (T_h \tilde{V}_{h+1}) (s_{h,r}, a_{h,r}) / \hat{\sigma}_h^2(s_{h,r}, a_{h,r}) \right)
\]
\[\leq \mathbb{E}_{\pi_h} [\phi]^T \hat{\Lambda}_h^{-1} A_h + \mathbb{E}_{\pi_h} [\phi]^T \hat{\Lambda}_h^{-1} A_h \leq \hat{O}(H^2 d / \kappa)
\]

For the second term, it can be bounded similar to term (iii) in Lemma D.1 and for the first term we have the following:
\[\mathbb{E}_{\pi_h} [\phi]^T \hat{\Lambda}_h^{-1} A_h = \mathbb{E}_{\pi_h} [\phi]^T \hat{\Lambda}_h^{-1} \cdot \hat{\Lambda}_h \hat{\Lambda}_h^{-1} A_h \leq \mathbb{E}_{\pi_h}[\phi] \hat{\Lambda}_h^{-1} A_h \leq \hat{O}(\sqrt{d} \mathbb{E}_{\pi_h}[\phi] \hat{\Lambda}_h^{-1}
\]

where the first inequality uses Cauchy’s inequality, the second inequality uses \(\hat{\Lambda}_h\) is coordinate-wise positive (since we assume here \(\phi \geq 0\), the third inequality is identical to the analysis in Section C.2.2 and the fourth inequality is identical to the analysis in Section C.2.1 with \(\phi\) replaced by \(\mathbb{E}[\phi]\). Plug this back to equation 22 we finish the proof for the first part. For the second part, converting \(\hat{\Lambda}_h^{-1}\) to \(\hat{\Lambda}_h^*\) is identical to Section C.5. This finishes the proof. \(\square\)
E PROOF OF MINIMAX LOWER BOUND THEOREM 3.5

The proof follows the lower bound proof of Zanette et al. (2021). For completeness, we provide all the details in below.

E.1 CONSTRUCTION

Similar to the proof of [Zanette et al. (2021), Theorem 2], we construct a family of MDPs, each parameterized by a Boolean vector $u = (u_1, \ldots, u_H)$ with each $u_h \in \{-1, +1\}^{d-2}$ for $h \in [H]$. The MDPs share the same transition kernel and are only different in the reward observations.

**State space:** At each time step $h$, there are two states $S = \{+1, -1\}$.

**Action space:** The action space $A = \{-1, 0, +1\}^{d-2}$.

**Feature map:** The feature map $\phi : S \times A \mapsto \mathbb{R}^d$ is given by

$$
\phi(+1, a) = \left( \frac{s}{\sqrt{2d}} \right) \in \mathbb{R}^d, \quad \phi(-1, a) = \left( \frac{s}{\sqrt{2d}} \right) \in \mathbb{R}^d.
$$

The construction ensures the condition $\|\phi(s, a)\|_2 \leq 1$ for any $(s, a) \in S \times A$.

**Transition kernel:** The transition probability $P_h(s' | s, a)$ is independent of action $a$. In other words, the Markov decision process reduces to a homogeneous Markov chain with transition matrix

$$
P = \begin{pmatrix}
\frac{1}{2} & \frac{1}{2} \\
\frac{1}{2} & \frac{1}{2}
\end{pmatrix} \in \mathbb{R}^2.
$$

By letting

$$
\nu_h(+1) = \nu_h(-1) = \left( \frac{0}{\sqrt{2d}} \right) \in \mathbb{R}^d,
$$

we have $P_h(s' | s, a) = \langle \phi(s, a), \nu_h(s') \rangle$ to be a valid probability transition.

**Reward observations:** For any MDP $M_u$, at each times step $h$, the reward follows a Gaussian distribution with

$$
R_{u,h}(s, a) \sim \mathcal{N}\left( \frac{s}{\sqrt{6}} + \frac{\delta}{\sqrt{2d}} \langle a, u_h \rangle, 1 \right),
$$

where $\delta \in \left[ 0, \frac{1}{\sqrt{3d}} \right]$ determines to what extent the MDP models are different from each other. The mean reward function satisfies $r_{u,h}(s, a) = \langle \phi(s, a), \theta_{u,h} \rangle$ with

$$
\theta_{u,h} = \left( \frac{\delta u_h}{\sqrt{3}} \right) \in \mathbb{R}^d.
$$

**Offline data collection Scheme:** The dataset $D = \{(s_h^0, a_h^0, r_h^0, s_h^1) \}_{h \in [H]}$ consist of $K$ i.i.d. trajectories. All the trajectories initiate from uniform distribution. We take a behavior policy $\mu(\cdot | s)$ that is independent of state $s$. Let $\{e_1, e_2, \ldots, e_{d-2}\}$ be the canonical bases of $\mathbb{R}^{d-2}$ and $0_{d-2} \in \mathbb{R}^{d-2}$ be the zero vector. The behavior policy $\mu$ is set as

$$
\mu(e_j | s) = \frac{1}{d} \quad \text{for any } j \in [d-2] \quad \text{and} \quad \mu(0_{d-2} | s) = \frac{2}{d}.
$$

E.2 OVERVIEW OF PROOF

The proof of the theorem is based on Assouad’s method, where we first reduce the problem to binary hypothesis tests (Lemmas E.1 and E.2) and then connect the testing error to the uncertainty quantity in the upper bound (Lemma E.3).
Lemma E.1 (Reduction to testing). There exists a universal constant \( c_1 > 0 \) such that:

\[
\inf_{\pi} \max_{u \in \mathcal{U}} \mathbb{E}_u [V^*_u - V^\pi_u] \geq c_1 \delta \sqrt{dH} \min_{u, u' \in \mathcal{U} : D_H(u; u') = 1} \inf_{\psi} [\mathbb{P}_u(\psi \neq u) + \mathbb{P}_{u'}(\psi \neq u')],
\]

(23)

where \( \hat{\pi} \) denotes the output of any algorithm that maps from observations to an estimated policy. \( \psi \) is any test function for parameter \( u \) and \( D_H \) is the hamming distance.

Lemma E.2. There exists a universal constant \( c_2 > 0 \) such that when taking \( \delta := \frac{c_2 d}{\sqrt{K}} \), we have:

\[
\min_{u, u' \in \mathcal{U} : D_H(u; u') = 1} \inf_{\psi} [\mathbb{P}_u(\psi \neq u) + \mathbb{P}_{u'}(\psi \neq u')] \geq \frac{1}{2}.
\]

(24)

When \( K \geq d^3 \), \( \delta := \frac{c_2 d}{\sqrt{K}} \) ensures that \( \delta \leq 1/\sqrt{3d} \). Combining Lemmas E.1 and E.2 yields a lower bound:

\[
\inf_{\pi} \max_{u \in \mathcal{U}} \mathbb{E}_u [V^*_u - V^\pi_u] \geq c \frac{d \sqrt{dH}}{\sqrt{K}},
\]

(25)

where \( c > 0 \) is a universal constant. We then use the following Lemma E.3 to connect the above lower bound to the uncertainty term \( \sqrt{d} \cdot \sum_{h=1}^H \sqrt{\mathbb{E}_\pi_{\cdot \cdot} [\phi] \phi} (\Lambda_h^\ast)^{-1} \mathbb{E}_\pi_{\cdot \cdot} [\phi] \) for the chosen linear MDP instances class \( \mathcal{M} \).

Lemma E.3. There exists a universal constant \( c_3 > 0 \) such that for all \( M \in \mathcal{M}, \)

\[
\sum_{h=1}^H \sqrt{\mathbb{E}_\pi_{\cdot \cdot} [\phi] \phi} (\Lambda_h^\ast)^{-1} \mathbb{E}_\pi_{\cdot \cdot} [\phi] \leq c_3 \frac{dH}{\sqrt{K}}.
\]

(26)

Plugging inequality equation E.3 into the bound equation 25, we obtain the minimax lower bound equation 6 in the statement of theorem.

E.3 REDUCTION TO TESTING VIA ASSOUAD’S METHOD

Proof of Lemma E.1. For any index vector \( u = (u_1, \ldots, u_H) \in \mathcal{U} = \{-1, +1\}^{(d-2) \times H} \), the optimal policy for MDP instance \( M_u \) is simply

\[
\pi_h^\ast(\cdot) = u_h \quad \text{for } h \in [H].
\]

Similar to the proof of Lemma 9 in Zanette et al. (2021), we can show that the value suboptimality of policy \( \pi \) on MDP \( M_u \) is given by

\[
V^*_u - V^\pi_u = \frac{\delta}{\sqrt{2d}} \sum_{h=1}^H \|u_h - \mathbb{E}_\pi [a_h]\|_1.
\]

Define \( u^\pi = (u_1^\pi, \ldots, u_H^\pi) \) with \( u_h^\pi := \text{sign}(\mathbb{E}_\pi [a_h]) \), then the \( \ell_1 \)-norm is lower bounded as

\[
\|u_h - \mathbb{E}_\pi [a_h]\|_1 \geq D_H(u_h^\pi; u_h),
\]

where \( D_H(\cdot; \cdot) \) denotes the Hamming distance. It follows that

\[
V^*_u - V^\pi_u \geq \frac{\delta}{\sqrt{2d}} D_H(u^\pi; u).
\]

(27)

We then apply Assouad’s method (Lemma 2.12 in Sampson and Guttorp (1992)) and obtain that

\[
\inf_{u \in \mathcal{U}} \max_{u' \in \mathcal{U}} \mathbb{E}_u [D_H(\bar{u}; u)] \geq \frac{(d - 2)H}{2} \min_{u, u' \in \mathcal{U} : D_H(u; u') = 1} \inf_{\psi} [\mathbb{P}_u(\psi \neq u) + \mathbb{P}_{u'}(\psi \neq u')],
\]

(28)

where \( \psi \) is any test functions mapping from observations to \( \{u, u'\} \). Combining inequalities equation 27 and equation 28, we finish the proof. \( \square \)
E.4 Lower Bound on the Testing Error

Proof of Lemma E.2. The proof of Lemma E.2 is similar to that of Lemma 10 in Zanette et al. (2021). We first apply Theorem 2.12 in Sampson and Guttorp (1992) to lower bound the testing error using Kullback–Leibler divergence and obtain

$$\min_{u,u' \in U : D_H(u';u)=1} \inf_{\psi} \left[ \mathbb{P}_u(\psi \neq u) + \mathbb{P}_{u'}(\psi \neq u') \right] \geq 1 - \left( \frac{1}{2} \max_{u,u' \in U : D_H(u';u)=1} D_{KL}(Q_u \| Q_{u'}) \right)^{1/2}.$$  \hfill (29)

It only remains to estimate $D_{KL}(Q_u \| Q_{u'})$.

The probability density $Q_u$ takes the form

$$Q_u(\mathcal{D}) = \prod_{k=1}^{K} \xi_k(s_k^1) \prod_{h=1}^{H} \mu(a_k^h | s_k^h) \left[ R_{u,h}(s_k^h, a_k^h) \right] \mathbb{P}_u(s_k^h+1 | s_k^h, a_k^h)$$

where $\xi_k = \left[ \frac{1}{2} \frac{1}{2} \right]$ is the initial distribution. It follows that

$$D_{KL}(Q_u \| Q_{u'}) = \mathbb{E}_u \left[ \log(Q_u / Q_{u'}) \right]$$

$$= K \cdot \sum_{h=1}^{H} \mathbb{E}_u \left[ \log \left( \left[ R_{u,h}(s_k^h, a_k^h) \right] \left[ R_{u',h}(s_k^h, a_k^h) \right] \right) \right]$$

$$= \frac{K}{d} \sum_{j=1}^{d-2} D_{KL} \left( \mathcal{N} \left( \frac{a}{\sqrt{2d}}(e_j, u_h), 1 \right) \bigg\| \mathcal{N} \left( \frac{a}{\sqrt{2d}}(e_j, u'_h), 1 \right) \right).$$

If we take $\delta = \frac{\varepsilon d}{\sqrt{N}}$, then inequality equation 29 ensures inequality equation 24, as claimed in the statement of the lemma.

\hfill $\square$

E.5 Connection to the Uncertainty Term

Proof of Lemma E.3. We first calculate the explicit form of the inverse of variance-rescaled covariance matrix $\Lambda_h^{-p}$. For each time step $h \in [H]$, the value function $V_{u,h+1}^{*}$ takes the form

$$V_{u,h+1}^{*} = \mathbb{E}_\pi \cdot r_{u,h+1} + \left( \mathbb{P}_{h+1} V_{u,h+2}^{*} \right)(+1).$$

Since $(\mathbb{P}_{h+1} V_{u,h+2}^{*})(+1) = (\mathbb{P}_{h+1} V_{u,h+2}^{*})(-1)$ and $r_{u,h+1}(+1,a) - r_{u,h+1}(-1,a) = 2/\sqrt{6}$, we have

$$\text{Var}_{p_h}(V_{u,h+1}^{*})(+1,a) = \text{Var}_{p_h}(\mathbb{E}_\pi \cdot r_{u,h+1})(+1,a) = \frac{1}{6}.$$

Similarly,

$$\text{Var}_{p_h}(V_{u,h+1}^{*})(-1,a) = \text{Var}_{p_h}(V_{u,h+1}^{*})(+1,a) = \frac{1}{6}.$$

By routine calculation, we find that the population-level rescaled covariance matrix takes the form

$$\Lambda_h^{-p} = \frac{3K}{2} \left( \begin{bmatrix} \frac{1}{d} \mathbb{I}_{d-2} & -\frac{1}{d\sqrt{d}} \mathbb{I}_{(d-2) \times 2} \\ -\frac{d\sqrt{d}}{2(d-2)} \mathbb{I}_{d-2} & \frac{1}{d-1} \mathbb{I}_{(d-2) \times 2} \end{bmatrix} \right) \in \mathbb{R}^{d \times d}$$

for any $h \in [H]$. Applying Gaussian elimination on $\Lambda_h^{-p}$, we have

$$(\Lambda_h^{-p})^{-1} = \frac{2}{3K} \left( \begin{bmatrix} \frac{d^2}{2} \mathbb{I}_{d-2} + \frac{1}{d-2} \mathbb{I}_{(d-2) \times (d-2)} & -\frac{d\sqrt{d}}{2(d-2)} \mathbb{I}_{(d-2) \times 2} \\ -\frac{d\sqrt{d}}{2(d-2)} \mathbb{I}_{d-2} & \frac{1}{d-2} \left( \begin{bmatrix} d-1 & 1 \\ 1 & d-1 \end{bmatrix} \right) \end{bmatrix} \right).$$
For each time step $h \in [H]$, we have (by Jensen’s inequality)
\[
\sqrt{E_\pi, [\phi] (\Lambda_h^*)^{-1} E_\pi, [\phi]} \leq \frac{1}{2} \|\phi(+1, u_h)\|_{(\Lambda_h^*)^{-1}} + \frac{1}{2} \|\phi(-1, u_h)\|_{(\Lambda_h^*)^{-1}}.
\]
Recall that by our construction,
\[
\phi(+1, u_h) = \left( \frac{u_h \sqrt{2d}}{\sqrt{2}} \right) \in \mathbb{R}^d, \quad \phi(-1, u_h) = \left( \frac{u_h \sqrt{2d}}{1 \sqrt{2}} \right) \in \mathbb{R}^d.
\]
It follows that
\[
\|\phi(+1, u_h)\|_{(\Lambda_h^*)^{-1}} = \|\phi(-1, u_h)\|_{(\Lambda_h^*)^{-1}} = \frac{2}{3K} \left\{ \frac{d^2}{4} + \frac{d}{2(d-2)} \left( 1 - \frac{1}{d-2} d^2 u_h + \frac{d-1}{2(d-2)} \right) \right\}
\]
\[
\leq \frac{2}{3K} \left\{ \frac{d^2}{4} + \frac{d(d-1)^2}{4(d-2)} + \frac{1}{4} \right\} = \frac{2}{3K} \left\{ \frac{1}{2} + \frac{d - 1}{2(d-2)} \right\} \lesssim d^2 / K.
\]
Therefore,
\[
\sqrt{E_\pi, [\phi] (\Lambda_h^*)^{-1} E_\pi, [\phi]} \lesssim d / \sqrt{K}.
\]
Taking the summation over $h \in [H]$, we obtain the bound equation 26 as claimed in the lemma statement.

\[\square\]

F A NUMERICAL SIMULATION

F.1 A LINEAR MDP CONSTRUCTION

We consider a synthetic linear MDP example that is similar to Min et al. (2021) but with some modifications for the offline learning task. The MDP instance we use consists of $|S| = 2$ states and $|A| = 100$ actions, and feature dimension $d = 100$. We set $S = \{0, 1\}$ and $A = \{0, 1, \ldots, 99\}$ respectively. For each action $a \in \{0, 1, \ldots, 99\}$, we use binary encoding to obtain a vector $a \in \mathbb{R}^8$ using its binary representation (i.e. each coordinate is either 0 or 1). We interchangeably use $a$ and its vector representation for the ease of explanation. We first define $\delta(s, a) = \left\{ \begin{array}{ll} 1 & \text{if } 1\{s = 0\} = 1\{a = 0\} \\ 0 & \text{otherwise} \end{array} \right.$, then the non-stationary linear MDP is specified by the following configuration:

- Feature mapping:
\[
\phi(s, a) = \left( a^\top, \delta(s, a), 1 - \delta(s, a) \right) \in \mathbb{R}^{10}
\]

- The true measure $\nu_h$
\[
\nu_h(s) = \{0, \ldots, 0, (1 - s) \oplus a_h, s \oplus a_h\},
\]
where $\{a_h\}_{h \in [H]}$ is a sequence of integers taking values 0 or 1 and $\oplus$ is the standard XOR operator. We define $\theta_h = \{0, \ldots, 0, r, 1 - r\} \in \mathbb{R}^{10}$ with the choice of $r = 0.9$. The transition follows $P_h(s'|s, a) = \langle \phi(s, a), \nu_h(s') \rangle$ and the mean reward function $r_h(s, a) = \langle \phi(s, a), \theta_h \rangle$.

- Behavior policy: always choose action $a = 0$ with probability $p$, and other actions uniformly with probability $(1 - p)/99$. The initial distribution chooses $s = 0$ and $s = 1$ with equal probability 1/2. We use $p = 0.6$. 

33
Figure 1: Comparison between PEVI and VAPVI in the non-stationary linear MDP instance described above. In each figure, y-axis denotes suboptimality gap $v^* - v^\pi$, x-axis denotes number of episodes $K$. The problem horizons are fixed to be $H = 20, 30, 50$. The solid line denotes the average suboptimality gap over 50 trials and the error bar area is the corresponding standard deviation. The range of $K$ is from 5 to 1000.

F.2 Empirical Comparison Between PEVI and VAPVI on the Constructed Linear MDP

We compare Pessimistic Value Iteration (PEVI) in Jin et al. (2021b) and our VAPVI Algorithm 1 in Figure 1, with horizon to be $H = 20, 30, 50$. In addition, we add the non-pessimistic version for both algorithms, i.e. least-square value iteration (LSVI) and variance-aware value iteration (VAVI). The true optimal value $v^*$ is computed via value iteration using the underlying transition kernels. For the empirical validation of VAPVI, we do not split the data and, in particular, in all the methods we choose $\lambda = 0.01$ (instead of $\lambda = 1$ used in theory (Jin et al., 2021b) which causes over-regularization in the simulation).

We can observe VAPVI outperforms PEVI and the gap becomes larger when horizon $H$ increases. One main reason for this to happen is due to the bonus used in PEVI (Jin et al., 2021b)

$$O \left[ dH \cdot \left( \phi(\cdot, \cdot)^T \Sigma_h^{-1} \phi(\cdot, \cdot) \right)^{1/2} \right]$$

is overly pessimistic comparing to our

$$O \left[ \sqrt{d} \cdot \left( \phi(\cdot, \cdot)^T \Lambda_h^{-1} \phi(\cdot, \cdot) \right)^{1/2} \right]$$

when $H$ becomes larger and this could potentially make the learning less accurate. In addition, both non-pessimistic algorithms exhibit similar accuracy, and this is partially owing to our truncation scheme $	ilde{\sigma}_h(\cdot, \cdot)^2 \leftarrow \max \{ 1, \hat{\text{Var}}_h \hat{V}_{h+1}(\cdot, \cdot) \}$ so $	ilde{\sigma}_h(\cdot, \cdot)^2$ will just be 1 when the estimated variance is small. Lastly, variance-aware pessimism eventually outperforms non-pessimism algorithms when sample size is large and this might come from that the pessimistic bonus is estimated more accurately when more samples are collected.

G Some Missing Derivations and Discussions

G.1 Regarding Coverage Assumption

Now we discuss the feature coverage assumption. Indeed, even if Assumption 2.2 is not satisfied, we can still learn in the effective subspace of $\Sigma_h' := E_{\mu,h} \left[ \phi(s,a)\phi(s,a)^T \right]$. Concretely, since $\Sigma_h'$ is symmetric, by orthogonal decomposition we have $\Sigma_h' = Z_h \Lambda_h Z_h^T$, where $Z_h$ (can be estimated using the samples for practical purpose) consists of orthogonal basis and $\Lambda$ consists of eigenvalues of $\Sigma_h'$ in the diagonal. Suppose we do not have a full coverage, i.e.

$$\Lambda = \text{diag}[\lambda_1, \lambda_2, ..., \lambda_{d'}, 0, ..., 0] \quad \text{with} \quad d' < d,$$

then we can create transformed features $\phi'_h(s,a) = Z_h \cdot \phi_h(s,a)$, and then

$$E_{\mu,h} \left[ \phi'_h(s,a)\phi'_h(s,a)^T \right] = \Lambda = \text{diag}[\lambda_1, \lambda_2, ..., \lambda_{d'}, 0, ..., 0].$$

Then we can do learning w.r.t. the truncated features $\phi'_h|_{1:d'}$’s instead of the original $\phi$. It reduces to the weaker notion of $\kappa := \min_{h \in [H]} \{ \kappa_h : s.t. \kappa_h = \text{smallest positive eigenvalue at time } h \}$. 

34
G.2 Derivation of equation 5

When reducing Theorem 3.2, 3.3 to the tabular case, set $\phi(s, a) = 1_{s,a}$, $d = SA$, $\lambda = 0$, and recall by Assumption 3.4 (let’s assume $\pi^*$ is a deterministic policy as it always exists in tabular MDP) $C^* := \sup_{h,s,a} d_h^*(s, a)/d_H^*(s, a)$, then for Theorem 3.2

\[
\sqrt{d} \cdot \sum_{h=1}^{H} \mathbb{E}_{\pi^*} \left[ \left( \phi, \Lambda_h^{s-1} \phi \right) \right] = \sqrt{d} \cdot \sum_{h=1}^{H} \sum_{s,a} d_h^*(s, a) \sqrt{1_{s,a} \Lambda_h^{s-1} 1_{s,a}}
\]

\[
= \sqrt{SA} \cdot \sum_{h=1}^{H} \sum_{s,a} d_h^*(s, a) \sqrt{\frac{\text{Var}_{P_{\pi^*}}(V_{h+1}^*)}{n_{h,s,a}}} \quad \text{where } n_{h,s,a} := \sum_{t=1}^{K} 1[s_t, a_t = s, a]
\]

\[
\leq \sqrt{SA} \cdot \sum_{h=1}^{H} \sum_{s,a} d_h^*(s, a) \sqrt{\frac{\text{Var}_{P_{\pi^*}}(V_{h+1}^*)}{K \cdot d_H^*(s, a)}} \leq \sqrt{SAC^* / K} \cdot \sum_{h=1}^{H} \sum_{s,a} d_h^*(s, a) \text{Var}_{P_{\pi^*}}(V_{h+1}^*)
\]

\[
= \sqrt{SAC^* / K} \cdot \sum_{h=1}^{H} \sum_{s} \sqrt{d_h^*(s, \pi^*(s)) \text{Var}_{P_{\pi^*}(s)}(V_{h+1}^*)}
\]

\[
\leq \sqrt{SAC^* / K} \cdot \sum_{h=1}^{H} \sqrt{S \cdot \sum_{s} d_h^*(s, \pi^*(s)) \text{Var}_{P_{\pi^*}(s)}(V_{h+1}^*)}
\]

\[
\leq \sqrt{S^2 AC^* / K} \cdot \sqrt{H \sum_{h=1}^{H} \sum_{s} d_h^*(s, \pi^*(s)) \text{Var}_{P_{\pi^*}(s)}(V_{h+1}^*)}
\]

\[
\leq \sqrt{S^2 AC^* / K} \cdot \sqrt{H \sum_{h=1}^{H} \mathbb{E}_{\pi^*} \left[ \text{Var}_{P_{\pi^*}(\cdot)}(V_{h+1}^*) \right]} \leq \sqrt{H^3 S^2 AC^*/K}
\]

where the first inequality is by Chernoff bound and the last one is by Lemma 3.4. of Yin and Wang (2020) (Law of total variances). The rest of them are from Cauchy’s inequality. Similarly, for Theorem 3.3, we also have

\[
\sqrt{d} \cdot \sum_{h=1}^{H} \mathbb{E}_{\pi^*} [\phi]^\top \Lambda_h^{s-1} \mathbb{E}_{\pi^*} [\phi] = \sqrt{d} \cdot \sum_{h=1}^{H} \sqrt{\text{Vec}(d^*)^\top \Lambda_h^{s-1} \text{Vec}(d^*)}
\]

\[
= \sqrt{d} \cdot \sum_{h=1}^{H} \sqrt{\text{Vec}(d^*)^\top \text{diag} \left( \frac{\text{Var}_{P_{\pi^*}}(V_{h+1}^*)}{n_{h,s,a}} \right) \text{Vec}(d^*)}
\]

\[
= \sqrt{SA} \cdot \sum_{h=1}^{H} \sum_{s,a} d_h^*(s, a) \sqrt{\frac{\text{Var}_{P_{\pi^*}}(V_{h+1}^*)}{n_{h,s,a}}}
\]

\[
\leq \sqrt{SA} \cdot \sum_{h=1}^{H} \sum_{s,a} d_h^*(s, a) \sqrt{\frac{\text{Var}_{P_{\pi^*}}(V_{h+1}^*)}{K \cdot d_H^*(s, a)}}
\]

\[
\leq \sqrt{SAC^* / K} \cdot \sum_{h=1}^{H} \sum_{s,a} d_h^*(s, a) \text{Var}_{P_{\pi^*}}(V_{h+1}^*)
\]

\[
= \frac{\sqrt{SAC^* / K} \cdot \sum_{h=1}^{H} \sum_{s,a} d_h^*(s, a) \text{Var}_{P_{\pi^*}(s)}(V_{h+1}^*)}{\sqrt{H \sum_{h=1}^{H} \mathbb{E}_{\pi^*} \left[ \text{Var}_{P_{\pi^*}(\cdot)}(V_{h+1}^*) \right]}} \leq \frac{\sqrt{H^3 S^2 AC^*/K}}{K}
\]
H Auxiliary Lemmas

Lemma H.1 (Matrix McDiarmid inequality / Matrix Chernoff bound (Tropp, 2012)). Let $z_k, k = 1, \ldots, K$ be independent random vectors in $\mathbb{R}^d$, and let $H$ be a mapping that maps $K$ vectors to a $d \times d$ symmetric matrix. Assume there exists a sequence of fixed symmetric matrices $\{A_k\}_{k \in [K]}$ such that for $z_k, z_k'$ ranges over all possible values for each $k \in [K]$, it holds

$$(H(z_1, \ldots, z_K) - H(z_1, \ldots, z_k', \ldots, z_K))^2 \leq A_k^2.$$ Define $\sigma^2 := \| \sum_k A_k^2 \|$. Then for any $t > 0$,

$$\Pr \left\{ \| H(z_1, \ldots, z_K) - \mathbb{E} H(z_1, \ldots, z_K) \| \geq t \right\} \leq d \cdot \exp \left( \frac{-t^2}{8\sigma^2} \right).$$

Lemma H.2 (Hoeffding inequality for self-normalized martingales (Abbasi-Yadkori et al., 2011)). Let $\{\eta_t\}_{t=1}^\infty$ be a real-valued stochastic process. Let $\{F_t\}_{t=0}^\infty$ be a filtration, such that $\eta_t$ is $F_t$-measurable. Assume $\eta_t$ also satisfies $\eta_t$ given $F_{t-1}$ is zero-mean and $R$-subgaussian, i.e.

$$\forall \lambda \in \mathbb{R}, \quad \mathbb{E} [e^{\lambda \eta_t} | F_{t-1}] \leq e^{\lambda^2 R^2/2}$$

Let $\{x_t\}_{t=1}^\infty$ be an $\mathbb{R}^d$-valued stochastic process where $x_t$ is $F_{t-1}$ measurable and $\|x_t\| \leq L$. Let $A_t = \lambda d + \sum_{s=1}^t x_s x_s^T$. Then for any $\delta > 0$, with probability $1 - \delta$, for all $t > 0$,

$$\left\| \sum_{s=1}^t x_s \eta_s \right\|_{\Lambda_t^{-1}}^2 \leq 8 R^2 \cdot \frac{d}{2} \log \left( \frac{\lambda + tL}{\lambda \delta} \right).$$

Lemma H.3 (Bernstein inequality for self-normalized martingales (Zhou et al., 2021a)). Let $\{\eta_t\}_{t=1}^\infty$ be a real-valued stochastic process. Let $\{F_t\}_{t=0}^\infty$ be a filtration, such that $\eta_t$ is $F_t$-measurable. Assume $\eta_t$ also satisfies

$$|\eta_t| \leq R, \quad \mathbb{E} [\eta_t | F_{t-1}] = 0, \quad \mathbb{E} [\eta_t^2 | F_{t-1}] \leq \sigma^2.$$ Let $\{x_t\}_{t=1}^\infty$ be an $\mathbb{R}^d$-valued stochastic process where $x_t$ is $F_{t-1}$ measurable and $\|x_t\| \leq L$. Let $A_t = \lambda d + \sum_{s=1}^t x_s x_s^T$. Then for any $\delta > 0$, with probability $1 - \delta$, for all $t > 0$,

$$\left\| \sum_{s=1}^t x_s \eta_s \right\|_{\Lambda_t^{-1}} \leq 8 \sigma \sqrt{d \log \left( 1 + \frac{tL^2}{\lambda d} \right) \cdot \log \left( \frac{4t^2}{\delta} \right) + 4R \log \left( \frac{4t^2}{\delta} \right)}$$

Lemma H.4 (Converting the variance under the matrix norm). Let $\Lambda_1$ and $\Lambda_2 \in \mathbb{R}^{d \times d}$ are two positive semi-definite matrices. Then:

$$\| \Lambda_1^{-1} \| \leq \| \Lambda_2^{-1} \| + \| \Lambda_1^{-1} \| \cdot \| \Lambda_2^{-1} \| \cdot \| \Lambda_1 - \Lambda_2 \|$$

and

$$\| \phi \|_{\Lambda_1^{-1}} \leq \left[ 1 + \sqrt{\| \Lambda_2^{-1} \| \cdot \Lambda_2^{-1} \| \cdot \| \Lambda_1 - \Lambda_2 \|} \right] \cdot \| \phi \|_{\Lambda_2^{-1}}.$$ for all $\phi \in \mathbb{R}^d$.

Proof. For the first part, note

$$\| \Lambda_1^{-1} \| \leq \| \Lambda_2^{-1} \| + \| \Lambda_1^{-1} - \Lambda_2^{-1} \| \leq \| \Lambda_2^{-1} \| + \| \Lambda_2^{-1} \| \cdot \| \Lambda_1 - \Lambda_2 \| \cdot \| \Lambda_1^{-1} \|$$

For the second one,

$$\| \phi \|_{\Lambda_1^{-1}} = \sqrt{\phi^T \Lambda_1^{-1} \phi} = \sqrt{\phi^T (\Lambda_1^{-1} - \Lambda_2^{-1}) \phi + \phi^T \Lambda_2^{-1} \phi}$$

$$= \sqrt{\phi^T \Lambda_2^{-1/2} (\Lambda_2^{1/2} \Lambda_1^{-1} \Lambda_2^{1/2} - I + I) \Lambda_2^{-1/2} \phi} \leq \sqrt{\| \phi \|_{\Lambda_2^{-1}} \cdot \left( 1 + \| \Lambda_2^{1/2} \Lambda_1^{-1} \Lambda_2^{1/2} - I \| \right)} \cdot \| \phi \|_{\Lambda_2^{-1}}$$

$$\leq \left( 1 + \| \Lambda_2 \| \cdot \| \Lambda_1^{-1} \| \cdot \| \Lambda_2^{-1} \| \cdot \| \Lambda_1 - \Lambda_2 \| \right) \cdot \| \phi \|_{\Lambda_2^{-1}}$$

\[ \square \]
Lemma H.5 (Lemma H.4 of Min et al. (2021)). Let \( \phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d \) satisfy \( \| \phi(s, a) \| \leq C \) for all \( s, a \in \mathcal{S} \times \mathcal{A} \). For any \( K > 0, \lambda > 0 \), define \( \bar{G}_K = \sum_{k=1}^{K} \phi(s_k, a_k)\phi(s_k, a_k)\top + \lambda I_d \) where \( (s_k, a_k)'s \) are i.i.d samples from some distribution \( \nu \). Then with probability \( 1 - \delta \),
\[
\frac{\| \bar{G}_K - \mathbb{E}_{\nu} \left[ \frac{1}{K} \bar{\sum}_{\nu\cdot} \right] \|}{\sqrt{K}} \leq \frac{4\sqrt{2C^2}}{\sqrt{2d}} \left( \log \frac{2d}{\delta} \right)^{1/2}.
\]

Proof of Lemma H.5. For completeness, we provide the proof of Lemma H.5. Let \( x_k = \phi(s_k, a_k) \). Denote \( \Sigma_h \) as the matrix obtained by replacing the \( k \)-th vector \( x_k \) in \( \Sigma_h \) by \( \bar{x}_k \) and leaving the rest \( K - 1 \) vectors unchanged. Then
\[
\left( \frac{\Sigma_h \Sigma_h - \Sigma_h \Sigma_h}{K} \right)^2 = \left( \frac{x_k x_k\top - \bar{x}_k \bar{x}_k\top}{K} \right)^2 \leq \frac{1}{K^2} \left( 2x_k x_k\top x_k x_k\top + 2\bar{x}_k \bar{x}_k\top \bar{x}_k \bar{x}_k\top \right) \leq \frac{4C^4}{K^2} \mathbb{I} := A_k^2.
\]
Notice that \( \left\| \sum_{k=1}^{K} A_k^2 \right\| \leq \frac{4C^4}{K^2} \), by Lemma H.1 we have the result.

Lemma H.6 (Lemma H.5. of Min et al. (2021)). Let \( \phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d \) be a bounded function s.t. \( \| \phi \|_2 \leq C \). Define \( \bar{G}_K = \sum_{k=1}^{K} \phi(s_k, a_k)\phi(s_k, a_k)\top + \lambda I_d \) where \( (s_k, a_k)'s \) are i.i.d samples from some distribution \( \nu \). Let \( G = \mathbb{E}_\nu[\phi(s, a)\phi(s, a)\top] \). Then for any \( \delta \in (0, 1) \), if \( K \) satisfies
\[
K \geq \max \left\{ 512C^4 \| G^{-1} \|^2 \log \left( \frac{2d}{\delta} \right), 4\lambda \| G^{-1} \| \right\}.
\]
Then with probability at least 1 - \( \delta \), it holds simultaneously for all \( u \in \mathbb{R}^d \) that
\[
\| u \|_{G^{-1}} \leq \frac{2}{\sqrt{K}} \| u \|_{G^{-1}}.
\]

Lemma H.7 (Extended Value Difference (Section B.1 in Cai et al. (2020))). Let \( \pi = \{ \pi_h \}_{h=1}^H \) and \( \pi' = \{ \pi'_h \}_{h=1}^H \) be two arbitrary policies and let \( \{ \hat{Q}_h \}_{h=1}^H \) be any given \( Q \)-functions. Then define \( \hat{V}_h(s) := (\hat{Q}_h(s, \cdot), \pi_h(s, \cdot | s)) \) for all \( s \in \mathcal{S} \). Then for all \( s \in \mathcal{S} \),
\[
\hat{V}_1(s) - V_1'(s) = \sum_{h=1}^{H} \mathbb{E}_{\pi'}[\langle \hat{Q}_h(s, \cdot), \pi_h(s, \cdot | s) - \pi'_h(s, \cdot | s) \rangle | s = s] + \sum_{h=1}^{H} \mathbb{E}_{\pi'}[\langle \hat{Q}_h(s_h, a_h), (\mathcal{T}_h \hat{V}_{h+1})(s_h, a_h) - \hat{Q}_h(s_h, a_h) \rangle | s = s]
\]
where \( (\mathcal{T}_h V)(\cdot, \cdot) := r_h(\cdot, \cdot) + (P_h V)(\cdot, \cdot) \) for any \( V \in \mathbb{R}^S \).

Proof. Denote \( \hat{\xi}_h = \hat{Q}_h - \mathcal{T}_h \hat{V}_{h+1} \). For any \( h \in [H] \), we have
\[
\hat{V}_h - V_1'(h) = (\hat{Q}_h, \pi_h) - (\hat{Q}_h', \pi_h') = (\hat{Q}_h, \pi_h - \pi_h') + (\hat{Q}_h - \hat{Q}_h', \pi_h') = (\hat{Q}_h, \pi_h - \pi_h') + (P_h (\hat{V}_{h+1} - V_1'(h+1)) + \xi_h, \pi_h') = (\hat{Q}_h, \pi_h - \pi_h') + (P_h (\hat{V}_{h+1} - V_1'(h+1)) + \xi_h, \pi_h')
\]
recursively apply the above for \( \hat{V}_{h+1} - V_1'(h+1) \) and use the \( \mathbb{E}_{\pi'} \) notation (instead of the inner product of \( P_h, \pi'_h \)) we can finish the prove of this lemma.

Lemma H.8. Let \( \hat{\pi} = \{ \hat{\pi}_h \}_{h=1}^H \) and \( \hat{Q}_h(\cdot, \cdot) \) be the arbitrary policy and \( Q \)-function and also \( \hat{V}_h(s) := (\hat{Q}_h(s, \cdot), \hat{\pi}_h(\cdot | s)) \) for all \( s \in \mathcal{S} \) and \( \zeta_h(s, a) := \langle \mathcal{T}_h \hat{V}_{h+1}(s), a \rangle - \hat{Q}_h(s, a) \) (element-wise)
to be the Bellman update error. Then for any arbitrary $\pi$, we have
\[
V_1^\pi(s) - V_1^{\tilde{\pi}}(s) = \sum_{h=1}^{H} \mathbb{E}_\pi [\zeta_h(s_h, a_h) \mid s_1 = s] - \sum_{h=1}^{H} \mathbb{E}_{\tilde{\pi}} [\zeta_h(s_h, a_h) \mid s_1 = s] + \sum_{h=1}^{H} \mathbb{E}_\pi \left[ \langle \tilde{Q}_h(s_h, \cdot), \pi_h(\cdot|s_h) - \tilde{\pi}_h(\cdot|s_h) \rangle \mid s_1 = s \right]
\]
where the expectation are taken over $s_h, a_h$.

**Proof.** Note the gap can be rewritten as
\[
V_1^\pi(s) - V_1^{\tilde{\pi}}(s) = V_1^\pi(s) - \tilde{V}_1(s) - \tilde{V}_1(s) - V_1^{\tilde{\pi}}(s).
\]
By Lemma H.7 with $\pi = \tilde{\pi}, \pi' = \pi$, we directly have
\[
V_1^\pi(s) - \tilde{V}_1(s) = \sum_{h=1}^{H} \mathbb{E}_\pi [\zeta_h(s_h, a_h) \mid s_1 = s] + \sum_{h=1}^{H} \mathbb{E}_\pi \left[ \langle \tilde{Q}_h(s_h, \cdot), \pi_h(\cdot|s_h) - \tilde{\pi}_h(\cdot|s_h) \rangle \mid s_1 = s \right].
\]
Next apply Lemma H.7 again with $\pi = \pi' = \tilde{\pi}$, we directly have
\[
\tilde{V}_1(s) - V_1^{\tilde{\pi}}(s) = - \sum_{h=1}^{H} \mathbb{E}_{\tilde{\pi}} [\zeta_h(s_h, a_h) \mid s_1 = s].
\]
Combine the above two results we prove the stated result. \qed

**Lemma H.9.** For a linear MDP, for any $0 \leq V(\cdot) \leq H$, there exists a $w_h \in \mathbb{R}^d$ s.t. $T_h V = \langle \phi, w_h \rangle$ and $\|w_h\|_2 \leq 2H\sqrt{d}$ for all $h \in [H]$. Here $T_h(V)(s, a) = r_h(s, a) + (P_h V)(s, a)$. Similarly, for any $\pi$, there exists $w_h^{\pi} \in \mathbb{R}^d$, such that $Q_h^{\pi} = \langle \phi, w_h^{\pi} \rangle$ with $\|w_h^{\pi}\|_2 \leq 2(H - h + 1)\sqrt{d}$.

**Proof.** By definition,
\[
T_h V = r_h + (P_h V) = \langle \phi, \theta_h \rangle + \langle \phi, \int_S V(s) d\nu_h(s) \rangle
\]
\[
= \theta_h + \int_S V(s) d\nu_h(s),
\]
therefore $\|w_h\|_2 \leq \|\theta_h\|_2 + H \cdot \|\nu_h(S)\| \leq 1 + H\sqrt{d} \leq 2H\sqrt{d}$. The proof of the second part is similar by backward induction and the fact $V_1^\pi \leq H - h + 1$ for any $\pi$. \qed

**Lemma H.10.** For any pessimistic bonus design $\Gamma_h$, suppose $K > \max\{M_1, M_2, M_3, M_4\}$, then with probability $1 - \delta$, Algorithm 1 yields
\[
\left\| \widehat{T}_h \widehat{V}_{h+1} - \widehat{T}_h \widehat{V}_{h+1} \right\|_{\infty} \leq \tilde{O}\left( \frac{H^2\sqrt{d/K}}{\sqrt{K}} \right)
\]

38
Proof of Lemma H.10. Suppose $w_h$ is the coefficient corresponding to the $T_h \tilde{V}_{h+1}$ (such $w_h$ exists by Lemma H.9), i.e. $T_h \tilde{V}_{h+1} = \phi^\top w_h$, and recall $(T_h \tilde{V}_{h+1}) (s, a) = \phi(s, a)^\top (w_h - \tilde{w}_h)$, then:

$$
\left( T_h \tilde{V}_{h+1} \right) (s, a) - (T_h \tilde{V}_{h+1}) (s, a) = \phi(s, a)^\top (w_h - \tilde{w}_h)
$$

$$
= \phi(s, a)^\top w_h - \phi(s, a)^\top \tilde{w}_h
= \phi(s, a)^\top \tilde{w}_h
= \phi(s, a)^\top \tilde{w}_h
\nonumber
$$

For term (i), it is bounded by $2\lambda \delta^{\frac{3}{2}/\kappa}$ with probability $1 - \delta$ by Lemma C.2.

For term (ii), by Cauchy inequality it is bounded by

$$
\|\phi(s, a)\|_{\tilde{\Lambda}_h}^{-1} \cdot \left\| \sum_{\tau = 1}^{K} \phi(s_h^\tau, a_h^\tau) \cdot (r_h + \tilde{V}_{h+1} (s_h^\tau+1) - (T_h \tilde{V}_{h+1}) (s_h^\tau, a_h^\tau) / \tilde{\sigma}_h^2 (s_h^\tau, a_h^\tau) \right\|_{\tilde{\Lambda}_h^{-1}}
\leq \frac{2H}{\sqrt{\kappa R}} \left\| \sum_{\tau = 1}^{K} \phi(s_h^\tau, a_h^\tau) \cdot (r_h + \tilde{V}_{h+1} (s_h^\tau+1) - (T_h \tilde{V}_{h+1}) (s_h^\tau, a_h^\tau) / \tilde{\sigma}_h^2 (s_h^\tau, a_h^\tau) \right\|_{\tilde{\Lambda}_h^{-1}}
\leq \frac{2H}{\sqrt{\kappa R}} \cdot \tilde{O} (H^2 \sqrt{d / \kappa} / \sqrt{K}),
\nonumber
$$

where the first inequality is by Lemma H.6 (with $\phi' = \phi / \tilde{\sigma}_h$ and $\|\phi / \tilde{\sigma}_h\| \leq \|\phi\| \leq 1 := C$) and the third inequality uses $\sqrt{a^\top \cdot A \cdot a} \leq \sqrt{\|a\|_2 \|A\|_2} \|a\|_2 = \|a\|_2 \sqrt{\|A\|_2}$ with $a$ to be either $\phi$ or $w_h$. Moreover, $\lambda_{\min}(\tilde{\Lambda}_h^\top) \geq \kappa / \max_{h, s, a} \tilde{\sigma}_h(s, a)^2 \geq \kappa / H^2$ implies $\left\| (\tilde{\Lambda}_h^\top)^{-1} \right\| \leq 2 / \kappa$. The second inequality comes from Lemma H.2 with $R = H$ since $|\eta_r| = |r_h + \tilde{V}_{h+1} (s_h^\tau+1) - (T_h \tilde{V}_{h+1}) (s_h^\tau, a_h^\tau) / \tilde{\sigma}_h(s_h^\tau, a_h^\tau)| \leq H$ and $|x_r| = |\phi(s_h^\tau, a_h^\tau) / \tilde{\sigma}_h(s_h^\tau, a_h^\tau)| \leq 1$.

The final result is obtained by absorbing the term (i) via the condition $K >\max\{\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4\}$. \hfill \Box

Lemma H.11. Suppose random variables $\|X\|_\infty \leq 2H$, $\|Y\|_\infty \leq 2H$, then

$$
|\text{Var}(X) - \text{Var}(Y)| \leq 8H \cdot \|X - Y\|_\infty .
$$

Proof of Lemma H.11.

$$
|\text{Var}(X) - \text{Var}(Y)| = |E[X^2] - E[Y^2] - (E[X^2] - E[Y^2])| = |E[(X + Y)(X - Y)] - (E[X + Y])(E[X - Y])| \nonumber
\leq E[|X + Y| \cdot |X - Y|] + 4H \cdot \|X - Y\|_\infty \nonumber
\leq 4H \mathbb{E}[|X - Y|] + 4H \cdot \|X - Y\|_\infty = 8H \cdot \|X - Y\|_\infty .
\nonumber
$$

\hfill \Box