Exploring Differential Privacy in Reinforcement Learning: Towards Optimal Regret Bounds in Linear MDPs

Sharan Sahu
Department of Statistics and Data Science
Cornell University
Ithaca, NY 14853
ss4329@cornell.edu

Abstract

Motivated by the recent adoption of reinforcement learning (RL) in personalized decision making that relies on using users' sensitive and private information, we study regret minimization in the episodic inhomogeneous linear Markov Decision Process (MDP) setting where the transition probabilities and reward functions are linear with respect to some feature mapping $\boldsymbol{\phi}(s, a)$ under the constraints of differential privacy (DP) and more specifically, a relaxation of DP that is compatible with online-learning settings called joint differential privacy (JDP). Prior work due to [Luyo et al. 2021,] in this setting achieves a rate of $\tilde{O}(\sqrt{d^3H^4K} + H^{11/5}d^{8/5}K^{3/5}/\epsilon^{2/5})$ and was subsequently improved to $\tilde{O}(\sqrt{d^3H^4K} + H^3d^{5/4}K^{1/2}/\epsilon^{1/2})$ by [Ngo et al. 2022,]. This bounds rely on $\tilde{O}(\sqrt{d^3H^4K})$ dependence, the cost of non-private learning, that arises from the regret achieves by LSVI-UCB [Jin et al. 2020,]. Recently, [He et al. 2023,] proposed LSVI-UCB^{++}, a minimax optimal algorithm that achieves regret $\tilde{O}(HD\sqrt{T})$ for the episodic inhomogeneous linear MDP setting using weighted ridge regression and upper confidence value iteration with a Bernstein-type exploration bonus. Additionally, prior work primarily utilized Hoeffding-type bounds, which are easier to use in analysis but result in suboptimal regret bounds. [Qiao and Wang 2024,] advanced this area by applying Bernstein-type bounds to more effectively control regret for linear MDPs in the offline setting. Inspired by these works, we design an RL algorithm with differential privacy guarantees in the linear MDP setting by privatizing LSVI-UCB^{++}, utilizing the techniques found in [Qiao and Wang 2024,]. This algorithm achieves regret $\tilde{O}\left(d\sqrt{H^3K} + H^{18/4}d^{7/6}K^{1/2}/\epsilon\right)$ which surpasses previous state-of-the-art algorithms for linear MDPs. We also find that theory and simulation suggest that the privacy guarantee comes at (almost) no drop in utility compared to the non-private counterpart.

1 Introduction

Reinforcement Learning (RL) has started gaining traction in settings involving personalized decision-making such as precision medicine [Yazzourh et al. 2024, Liu et al. 2022,], user experience adaption [Khamaj and Ali 2024,], recommender systems [Afsar et al. 2022,], and autonomous driving [Sallab et al. 2017,]. In such settings, agents learn reasonable policies by learning from potentially private and sensitive user feedback and data. For example, imagine a health-focused mobile application designed to help users adopt a healthier lifestyle by recommending daily activities tailored to their needs and goals. This agent learns an optimal policy by observing user feedback, such as completion rates of recommended activities, user satisfaction ratings, and other behavioral signals. This process

inherently involves sensitive data—information that users may consider private, such as their age, weight, location, health habits, and physical activity levels. [Hartley et al. 2023,] recently showed that patient information can be memorized by agents even when it occurs on a single training data sample within the dataset.

To safeguard users' privacy, it's essential to incorporate privacy-preserving mechanisms into the RL framework. Differential Privacy (DP) [Dwork et al. 2006b,] has emerged as a rigorous mathematical notion of privacy in algorithms. The guarantee of a differentially private RL algorithm is that its behavior hardly changes when a single individual joins or leaves the dataset. It turns out for problems in this RL setting, the standard definition of DP is too stringent since it necessarily implies that in a setting where a user trusts an central agency with sensitive information in exchange for a service or recommendation, none of the agent's recommendations could reveal information about the user. Completely eliminating information about a user's data would make it impossible for the agent to make useful recommendations or actions.

We rely on a Joint Differential Privacy (JDP), a relaxed notion of DP [Kearns et al. 2015,]. JDP requires that if any single user changes their data, the information observed by all the other users cannot change substantially and has been adapted in the context of differentially private contextual bandits [Shariff and Sheffet 2018,]. There has been a line of literature that attempts to tackle incorporating JDP into RL algorithms in the linear MDP setting. The first work that we are aware of is [Luyo et al. 2021,] who privatize LSVI-UCB [Jin et al. 2020,] to get a regret bound of $\tilde{O}(\sqrt{d^3H^4K} + H^{11/5}d^{8/5}K^{3/5}/\epsilon^{2/5})$. This was subsequently improved to $\tilde{O}(\sqrt{d^3H^4K} + H^3d^{5/4}K^{1/2}/\epsilon^{1/2})$ by Ngo et al. [2022] through more refined analysis. Both these works rely on self-normalizing martingale concentration bounds, notably Azuma-Hoeffding, for their regret analysis. This allows for the analysis to be simple but results in suboptimal regret bounds.

Recently, [Qiao and Wang 2024,] were able to apply self-normalized Bernstein-type martingale bounds with sharper analysis to more effectively control regret for linear MDPs in the offline setting. Additionally, [He et al. 2023,] proposed LSVI-UCB^{++}, a minimax optimal algorithm that achieves regret $\tilde{O}(HD\sqrt{T})$ for the episodic inhomogeneous linear MDP setting using weighted ridge regression and upper confidence value iteration with a Bernstein-type exploration bonus which improved on LSVI-UCB.

Our contributions. Inspired by these works, we design an RL algorithm with differential privacy guarantees in the linear MDP setting by privatizing LSVI-UCB^{++}, utilizing the techniques found in [Qiao and Wang 2024,].

- We propose the DP-LSVI-UCB^{++} algorithm, which achieves a regret bound of $\tilde{O}\left(d\sqrt{H^3K} + H^{18/4}d^{7/6}K^{1/2}/\epsilon\right)$, surpassing the previous state-of-the-art bounds for linear MDPs under JDP constraints.
- LSVI-UCB^{++} framework, we integrate private mechanisms such as Gaussian noise and Gaussian Orthogonal Ensemble (GOE) perturbations of Gram matrices, enabling the preservation of privacy while maintaining strong utility guarantees.
- Our analysis employs Bernstein-type martingale concentration inequalities, unlike prior approaches relying on Hoeffding-type bounds, leading to tighter and more efficient regret guarantees.
- We provide empirical simulations that demonstrate the effectiveness of DP-LSVI-UCB^{++}, showcasing (almost) no drop in utility compared to its non-private counterpart across various privacy budgets.

1.1 Related work

Tabular MDPs: The intersection of DP and RL has been explored within the context of tabular MDPs. In these cases, DP is often achieved through privatization of visitation counts, which ensures that sensitive trajectory data remains protected. Under the constraint of JDP, [Vietri et al. 2020,] designed PUCB by privatizing UBEV [Dann et al. 2017,], and [Chowdhury and Zhou 2022,] devised Private-UCB-VI by privatizing UCBVI (with bonus 1) [Azar et al. 2017,], an algorithm with a minimax optimal regret bound in the tabular MDP setting. However, these works primarily utilized Hoeffding-type bounds, which are easier to use in analysis but result in suboptimal regret bounds.

[Qiao and Wang 2024,] advanced this area by applying Bernstein-type bounds to more effectively control regret, and Qiao and Wang [2023] designed DP-UCBVI by privatizing UCBVI with bonus 2 [Azar et al. 2017,].

Linear Mixture MDPs: There has been some work in the Linear Mixture MDP setting. Under JDP, [Luyo et al. 2021,] devised JDP-UCRL-VTR by privatizing UCRL-VTR [Ayoub et al. 2020,] with a regret bound $\tilde{O}(\sqrt{d^2H^4K} + H^{9/4}d^{3/4}K^{1/2}/\epsilon^{1/2})$ where K is the number of episodes. [Zhou 2022,] improved on this bound with Private-LinOpt-VI to guarantee JDP with a regret bound of $\tilde{O}(\sqrt{d^2H^4K} + H^{5/2}d^{7/4}K^{1/2}/\epsilon^{1/2})$.

Linear MDPs: There has also been some work in linear MDPs. Under JDP, [Luyo et al. 2021,] devised Privacy-Preserving LSVI-UCB Through Batching by privatizing LSVI-UCB [Jin et al. 2020,], and achieved a regret bound of $\tilde{O}(\sqrt{d^3H^4K}+H^{11/5}d^{8/5}K^{3/5}/\epsilon^{2/5})$ by utilizing standard differential private techniques such as the binary tree mechanism [Shariff and Sheffet 2018, Dwork et al. 2010, Chan et al. 2011,] and Gaussian mechanism [Dwork and Roth 2014,]. [Ngo et al. 2022,] improved on this bound, achieving $\tilde{O}(\sqrt{d^3H^4K} + H^3d^{5/4}K^{1/2}/\epsilon^{1/2})$ regret by utilizing an adaptive batching schedule to reduce the number of policy updates from polynomial in K to $O(\log(K))$.

2 Problem Setup

Markov Decision Process. We will work with the episodic inhomogeneous finite horizon MDP $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \{\mathbb{P}_h\}_h, \{r_h\}_h\}$ where $\mathcal{S}, \mathcal{A}$ is the state and action space respectively, $H \in \mathbb{Z}$ is the length of each episode, $\mathbb{P}_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ and $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$ are the time-dependent transition probability and deterministic reward function. We assume that $\mathcal{S}$ is a measurable space with possibly infinite number of elements and $\mathcal{A}$ is a finite set. In this setting, the policy is time-dependent and we denote this $\pi = \{\pi_1, \cdots, \pi_H\}$ where $\pi_h(s)$ denotes the action the policy takes in state s at timestep h. With this, we define the time-dependent value function $V_h^\pi : \mathcal{S} \to \mathbb{R}$ as

$$V_h^\pi(s) = \mathbb{E}\left[\sum_{t=h}^{H} r_t(s_t, a_t) \mid s_h = s, a_t \sim \pi_t(s_t)\right]$$

for any $s \in \mathcal{S},\ h \in [H]$. Likewise, we can define the state-action function $Q_h^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ as

$$Q_h^\pi(s, a) = \mathbb{E}\left[\sum_{t=h}^{H} r_t(s_t, a_t) \mid s_h = s,\ a_h = a,\ a_t \sim \pi_t(s_t)\right]$$

for any $s, a \in \mathcal{S} \times \mathcal{A},\ h \in [H]$. Since we are working in a finite episode length and action space, we know that there exists an optimal policy π^* such that $V_h^*(s) = \sup_\pi V_h^\pi(s)$ for any $s \in \mathcal{S},\ h \in [H]$ with Bellman equations

$$\begin{aligned} Q_h^*(s, a) &= r_h(s, a) + \mathbb{P}_h V_{h+1}^*(s, a) \\ V_h^*(s, a) &= \max_{a \in \mathcal{A}} Q_h^*(s, a) \end{aligned}$$

where $\mathbb{P}_h V(s, a) = \mathbb{E}_{s' \sim \mathbb{P}_h}(\cdot|s, a) V(s')$. We measure the performance of online reinforcement learning algorithms by the regret. The regret of an algorithm is defined as

$$\text{Regret}(K) = \sum_{k=1}^{K} \left[V_1^*\left(s_1^k\right) - V_1^{\pi_k}\left(s_1^k\right)\right]$$

where s_1 is the initial state and π_k is the agent, both during episode k.

Linear MDP [Jin et al. 2020,]. A finite-horizon MDP $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \{\mathbb{P}_h\}_h, \{r_h\}_h\}$ is a linear MDP with known feature map $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ if for any $h \in [H]$, there exists $|\mathcal{S}|$ unknown d-dimensional measures $\boldsymbol{\mu}_h = (\mu_h(1), \cdots, \mu_h(|\mathcal{S}|)) \in \mathbb{R}^{d \times |\mathcal{S}|}$ and an unknown vector $\boldsymbol{\theta}_h \in \mathbb{R}^d$ such that for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have

$$\mathbb{P}_h(\cdot \mid s,a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{\mu}_h(\cdot) \rangle,\ r_h(s,a) = \langle \boldsymbol{\phi}(s,a), \boldsymbol{\theta}_h(s,a) \rangle$$

Without loss of generality, we assume that $||\boldsymbol{\phi}(s,a)||_2 < 1$ and $\max\left(||\boldsymbol{\mu}_h(\mathcal{S})||_2, ||\boldsymbol{\theta}_h||_2\right) \leq \sqrt{d}$ for all $s, a, h \in \mathcal{S} \times \mathcal{A} \times [H]$.

2.1 Differential Privacy

In this work, we are interested in providing a privacy-preserving RL algorithm that incorporates the rigorous notion of differential privacy (DP). We first revisit the definition of differential privacy

Definition 2.1 (Differential Privacy [Dwork et al. 2006a,])**.** *A randomized mechanism A satisfies (ϵ, δ)-differential privacy if for all neighboring datasets $\mathcal{U}, \mathcal{U}'$ that differ by one record and for all event E in the output range*

$$\mathbb{P}\left(A(\mathcal{U}) \in E\right) \leq e^{\epsilon}\mathbb{P}\left(A(\mathcal{U}') \in E\right) + \delta$$

When $\delta = 0$, we say that our mechanism satisfies ϵ-pure DP whereas for $\delta > 0$, we say our mechanism satisfies (ϵ, δ)-DP.

As we discussed in the introduction, standard DP is too stringent of a framework to work in for the RL setting. Thus, we use JDP as a relaxed but still strong notion of privacy

Definition 2.2 (Joint Differential Privacy [Kearns et al. 2015,])**.** *For any $\epsilon > 0$, a randomized mechanism $A : \mathcal{U} \to \mathcal{A}^{KH}$ is ϵ-joint differentially private if for any $k \in [K]$, any user sequences $\mathcal{U}, \mathcal{U}'$ differing on the k-th user and any $E \subset \mathcal{A}^{(K-1)H}$*

$$\mathbb{P}\left(A_{-k}(\mathcal{U}) \in E\right) \leq e^{\epsilon}\mathbb{P}\left(A_{-k}(\mathcal{U}') \in E\right)$$

where $A_{-k}(\mathcal{U}) \in E$ denotes the sequence of actions recommended to all users except user k belong to the set E.

While we state our main results in terms of JDP, we will also use zero-Concentrated DP (zCDP) as a tool in our analysis, since it enables cleaner analysis for privacy composition and the Gaussian mechanism.

Definition 2.3 (zCDP [Dwork and Rothblum 2016, Bun and Steinke 2016,])**.** *A randomized mechanism A satisfies ρ-Zero-Concentrated Differential Privacy (ρ-zCDP), if for all neighboring datasets $\mathcal{U}$, $\mathcal{U}'$ and all $\alpha \in (1, \infty)$,*

$$D_{\alpha}\left(A\left(\mathcal{U}\right) || \ A\left(\mathcal{U}'\right)\right) \leq \rho\alpha$$

where D_{α} is the Renyi-divergence [van Erven and Harremoës 2012,]

Any algorithm that satisfies ρ-zCDP also satisfies approximate-DP. The following proposition from [Bun and Steinke 2016,] shows how to do the mapping between zCDP and approximate-DP.

Lemma 2.1 (Converting zCDP to DP [Bun and Steinke 2016,])**.** *If mechanism A satisfies ρ-zCDP, then A satisfies $\left(\rho + 2\sqrt{\rho \log\left(1/\delta\right)}, \delta\right)$-DP.*

Another simple and important property of zCDP is that compositions of zCDP mechanisms is also zCDP and any post-processing will not affect the privacy guarantees.

Lemma 2.2 (Adaptive composition and Post processing of zCDP [Bun and Steinke 2016,])**.** *Let $A : \mathcal{X}^n \to \mathcal{Y}$ and $A' : \mathcal{X}^n \times \mathcal{Y} \to \mathcal{Z}$. Suppose A satisfies ρ-zCDP and A' satisfies ρ'-zCDP. Define $A'' : \mathcal{X}^n \to \mathcal{Z}$ to be $A''(x) = A'(x, A(x))$. Then, A'' is $(\rho + \rho')$-zCDP.*

To apply DP techniques to some mechanism, we must know the sensitivity of the function we want to release. Here we give the definition and the notation we use.

Definition 2.4 (l_2-sensitivity)**.** *Let $\mathcal{U} \sim \mathcal{U}'$ be neighboring datasets. Then the l_2-sensitivity of a function $f : \mathbb{N}^{\mathcal{X}} \to \mathbb{R}^d$ is*

$$\Delta\left(f\right) = \max_{\mathcal{U} \sim \mathcal{U}'} ||f\left(\mathcal{U}\right) - f\left(\mathcal{U}'\right)||_2$$

In our analysis, we use the Gaussian mechanism:

Lemma 2.3 (Privacy guarantee of Gaussian mechanism [Dwork and Roth 2014, Bun and Steinke 2016,]). *Let $f : \mathbb{N}^{\mathcal{X}} \to \mathbb{R}^d$ be an arbitrary d-dimensional function with l_2 sensitivity Δ_2. The Gaussian Mechanism $\mathcal{M}$ with noise level σ is given by*

$$\mathcal{M}(\mathcal{U}) = f(\mathcal{U}) + \mathcal{N}\left(0, \sigma^2 I_d\right)$$

For any $\rho > 0$, a Gaussian Mechanism with noise parameter $\sigma^2 = \frac{\Delta_2^2}{2\rho}$ is ρ-zCDP. Additionally, for all $0 < \delta, \epsilon < 1$, a Gaussian Mechanism with noise parameter $\sigma = \frac{\Delta_2}{\epsilon}\sqrt{2\log\left(\frac{1.25}{\delta}\right)}$ satisfies (ϵ, δ)-DP.

Lastly, we use the following lemma to conclude that our algorithm is indeed joint differentially private

Lemma 2.4 (Billboard lemma [Hsu et al. 2013,]). *Suppose that a randomized mechanism $A : \mathcal{X}^n \to \mathcal{Y}$ is (ϵ, δ)-differentially private. Let $U \in \mathcal{U}$ be a dataset containing n users. Then, consider any set of functions $f_i : \mathcal{U}_i \times \mathcal{Y} \to \mathcal{Y}_i$ for $i \in [n]$ where $\mathcal{U}_i$ is the portion of the dataset containing user i's data. Then, the composition $\{f_i(\Pi_i(U), A(U))\}_{i\in[n]}$ is (ϵ, δ)-JDP where $\Pi : \mathcal{U} \to \mathcal{U}_i$ is the canonical projection to the i-th user's data.*

3 Main results

We now introduce our RL algorithm for linear MDPs with a JDP guarantee. We will first revisit the non-private version of LSVI-UCB^{++} proposed by [He et al. 2023,] and then we will propose our algorithm along with the techniques used to privatize LSVI-UCB^{++} with a desirable privacy-accuracy tradeoff.

LSVI-UCB^{++}. To estimate the parameter μ_h in linear MDPs, the LSVI-UCB^{++} algorithm employs a weighted ridge regression approach:

$$\Lambda_{k,h} = \lambda I + \sum_{i=1}^{k-1} \bar{\sigma}_{i,h}^{-2} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\top,$$

$$\widehat{w}_{k,h} = (\Lambda_{k,h})^{-1} \sum_{i=1}^{k-1} \bar{\sigma}_{i,h}^{-2} \phi(s_h^i, a_h^i)\widehat{V}_{k,h+1}(s_{h+1}^i),$$

$$\breve{w}_{k,h} = (\Lambda_{k,h})^{-1} \sum_{i=1}^{k-1} \bar{\sigma}_{i,h}^{-2} \phi(s_h^i, a_h^i)\breve{V}_{k,h+1}(s_{h+1}^i),$$

$$\overline{w}_{k,h} = \tilde{\Lambda}_{k,h}^{-1} \sum_{i=1}^{k-1} \bar{\sigma}_{i,h}^{-2} \phi\left(s_h^i, a_h^i\right) \widehat{V}_{k,h+1}(s_{h+1}^i)^2$$

where $\bar{\sigma}_{i,h}$ represents the variance of the optimal value function and is updated iteratively. This weighting by variance improves estimation accuracy by incorporating information about uncertainty. The optimistic and pessimistic value functions are updated as:

$$\widehat{Q}_{k,h}(s,a) = r_h(s,a) + \widehat{w}_{k,h}^\top \phi(s,a) + \widehat{\beta}\|\phi(s,a)\|_{\Lambda_{k,h}^{-1}},$$

$$\breve{Q}_{k,h}(s,a) = r_h(s,a) + \breve{w}_{k,h}^\top \phi(s,a) - \breve{\beta}\|\phi(s,a)\|_{\Lambda_{k,h}^{-1}},$$

with the corresponding state-value functions:

$$\widehat{V}_{k,h}(s) = \max_{a\in\mathcal{A}} \widehat{Q}_{k,h}(s,a), \quad \breve{V}_{k,h}(s) = \max_{a\in\mathcal{A}} \breve{Q}_{k,h}(s,a).$$

Here, $\widehat{\beta}$ and $\breve{\beta}$ determine the exploration bonuses, designed using Bernstein-type bounds. To ensure that the variance used for weighting in the ridge regression remains stable and avoids underestimation, we define a regularized variance for weighing:

$$\bar{\sigma}_{k,h} = \max\{\sigma_{k,h}, H, 2d^3H^2\|\phi(s_k^h, a_k^h)\|^{1/2}_{\Lambda_{k,h}^{-1}}\}$$

where the estimated variance $\sigma_{k,h}$ of the state-value function is

$$\sigma_{k,h} = \sqrt{\overline{\mathbb{V}}_{k,h}\widehat{V}_{k,h+1}(s_h^k, a_h^k) + E_{k,h} + D_{k,h} + H}$$

where we estimate the variance itself as

$$\overline{\mathbb{V}}_{k,h}\widehat{V}_{k,h+1}(s_h^k, a_h^k) = \left[\overline{w}_{k,h}^\top \phi\left(s_h^k, a_h^k\right)\right]_{[0,H^2]} - \left[\widehat{w}_{k,h}^\top \phi\left(s_h^k, a_h^k\right)\right]^2_{[0,H]}$$

Here, $E_{k,h}$ is the error between the estimated variance and the true variance of $V_{k,h+1}$, and $D_{k,h}$ is the error between the variance of $V_{k,h+1}$ and the variance of the optimal value function V_h^* [1]. When privatizing LSVI-UCB^{++}, we aim to privatize the individual statistics involved in making our final estimate $\widehat{w}_{k,h}$.

Private Model Components. In order to ensure differential privacy, the technique that we commonly employ in differential privacy is to cleverly add noise such that we achieve ρ-zCDP, but we also have utility of the specific statistics i.e. the private statistic is close to the non-private statistic with high probability. We add independent Gaussian noise to the $4HK$ statistics in DP-LSVI-UCB^{++} (Algorithm 1). Then, by the adaptive composition of zCDP (Lemma 2.2), it suffices to ensure that each statistic is ρ_0-zCDP where $\rho_0 = \frac{\rho}{4HK}$. In particular, in DP-LSVI-UCB^{++}, we utilize $\phi_1, \phi_2, \phi_3, K_1$ to denote the noise that we add. For all ϕ_i, we simply utilize the Gaussian Mechanism (Lemma 2.3). For K_1, we utilize a recent result by [Redberg and Wang 2021,] to release the Gram matrix using the GOE perturbations of the form $\frac{1}{\sqrt{2}}\left(Z + Z^\top\right)$. We also add $2\tilde{\lambda}_\Lambda$ to ensure that K_1 remains positive definite as the noise added violates this condition which we require for invertibility. In past literature, many resort to using a binary tree mechanism for privatizing the Gram matrix by recursively partitioning and privatizing partial sums. We find that privatization through GOE is better suited for this setting as it directly exploits their symmetry, yielding tighter utility bounds for the same privacy guarantees. We now present DP-LSVI-UCB^{++} (Algorithm 1)

Algorithm 1 DP-LSVI-UCB^{++}

Require: Confidence radius $\widehat{\beta}$, $\breve{\beta}$, $\tilde{\beta}$, Budget for zCDP ρ, Failure probability δ

Set $\rho_0 \leftarrow \frac{\rho}{4HK}$. Sample $\phi_1, \phi_2 \sim \mathcal{N}\left(0, \frac{2H^2}{\rho_0} I_d\right)$, $\phi_3 \sim \mathcal{N}\left(0, \frac{2H^4}{\rho_0} I_d\right)$, $K_1 \leftarrow \frac{1}{\sqrt{2}}\left(Z + Z^\top\right)$ where $Z_{i,j} \sim \mathcal{N}\left(0, \frac{1}{4\rho_0}\right)$, $\tilde{\lambda}_\Lambda = O\left(\sqrt{\frac{dHK}{\rho}}\right)$. Initialize $k_{\text{last}} = 0$ and for each stage $h \in [H]$, set $\tilde{\Lambda}_{0,h}, \tilde{\Lambda}_{1,h} \leftarrow 2\tilde{\lambda}_\Lambda I$

For each stage $h \in [H]$ and state-action $(s,a) \in S \times A$, set $\widehat{Q}_{0,h}(s,a) \leftarrow H$, $\breve{Q}_{0,h}(s,a) \leftarrow 0$

for episodes $k = 1, \ldots, K$ **do**

  Receive the initial state s_k^1

  for stage $h = H, \ldots, 1$ **do**

    $\tilde{\widehat{w}}_{k,h} \leftarrow \tilde{\Lambda}_{k,h}^{-1}\left[\sum_{i=1}^{k-1} \tilde{\widehat{\sigma}}_{i,h}^{-2} \phi\left(s_h^i, a_h^i\right) \tilde{\widehat{V}}_{k,h+1}(s_{h+1}^i) + \phi_1\right]$

    $\tilde{\breve{w}}_{k,h} \leftarrow \tilde{\Lambda}_{k,h}^{-1}\left[\sum_{i=1}^{k-1} \tilde{\widehat{\sigma}}_{i,h}^{-2} \phi\left(s_h^i, a_h^i\right) \tilde{\breve{V}}_{k,h+1}(s_{h+1}^i) + \phi_2\right]$

    $\tilde{\overline{w}}_{k,h} \leftarrow \tilde{\Lambda}_{k,h}^{-1}\left[\sum_{i=1}^{k-1} \tilde{\widehat{\sigma}}_{i,h}^{-2} \phi\left(s_h^i, a_h^i\right) \tilde{\widehat{V}}_{k,h+1}(s_{h+1}^i)^2 + \phi_3\right]$

    $\overline{\mathbb{V}}_{k,h}\tilde{\widehat{V}}_{k,h+1}(s_h^k, a_h^k) \leftarrow \left[\tilde{\overline{w}}_{k,h}^\top \phi\left(s_h^k, a_h^k\right)\right]_{[0,H^2]} - \left[\tilde{\widehat{w}}_{k,h}^\top \phi\left(s_h^k, a_h^k\right)\right]^2_{[0,H]}$

    if there exists a stage $h' \in [H]$ such that $\det(\tilde{\Lambda}_{k,h'}) \geq 2\det(\tilde{\Lambda}_{k_{\text{last}},h'})$ **then**

      $\tilde{\widehat{Q}}_{k,h}(s,a) \leftarrow \min\left\{r_h(s,a) + \tilde{\widehat{w}}_{k,h}^\top \phi(s,a) + \widehat{\beta}||\phi(s,a)||_{\tilde{\Lambda}_{k,h}^{-1}}, \tilde{\widehat{Q}}_{k-1,h}(s,a), H\right\}$

      $\tilde{\breve{Q}}_{k,h}(s,a) \leftarrow \min\left\{r_h(s,a) + \tilde{\breve{w}}_{k,h}^\top \phi(s,a) + \breve{\beta}||\phi(s,a)||_{\tilde{\Lambda}_{k,h}^{-1}}, \tilde{\breve{Q}}_{k-1,h}(s,a), 0\right\}$

      $k_{\text{last}} \leftarrow k$

    else

      $\tilde{\widehat{Q}}_{k,h}(s,a) = \tilde{\widehat{Q}}_{k-1,h}(s,a)$

[1] For more details about the LSVI-UCB^{++}, refer to [He et al. 2023,]

$\check{\tilde{Q}}_{k,h}(s,a) = \check{\tilde{Q}}_{k-1,h}(s,a)$
end if
$\hat{\tilde{V}}_{k,h}(s) = \max_{a\in\mathcal{A}} \hat{\tilde{Q}}_{k,h}(s,a)$
$\check{\tilde{V}}_{k,h}(s) = \max_{a\in\mathcal{A}} \check{\tilde{Q}}_{k,h}(s,a)$
$\tilde{\pi}_{k,h}(s) = \text{argmax}_{a\in\mathcal{A}} \hat{\tilde{Q}}_{k,h}(s,a)$
end for
for stage $h = 1, \ldots, H$ **do**
Take action $a_k^h \leftarrow \arg\max_a Q_{k,h}(s_k^h, a)$
$\tilde{\sigma}_{k,h} \leftarrow \sqrt{\overline{\mathbb{V}}_{k,h} \hat{\tilde{V}}_{k,h+1}(s_h^k, a_h^k) + E_{k,h} + D_{k,h} + H}$
$\tilde{\tilde{\sigma}}_{k,h} \leftarrow \max\{\tilde{\sigma}_{k,h}, H, 2d^3H^2\|\phi(s_k^h, a_k^h)\|^{1/2}_{\tilde{\Lambda}^{-1}_{k,h}}\}$
$\tilde{\Lambda}_{k+1,h} = \tilde{\Lambda}_{k,h} + \tilde{\tilde{\sigma}}^{-2}_{k,h}\phi(s_k^h, a_k^h)\phi(s_k^h, a_k^h)^\top + K_1$
Receive next state s_k^{h+1}
end for
end for

If one looks at the LSVI-UCB^{++} algorithm and compares it to our algorithm, one will notice that this algorithm is very similar except instead of using the raw statistics, we replace them with private ones as we described above. Since our algorithms are the same, most of the analysis carried out will be similar except that we will use the utility of the privatized statistics. We now present the privacy guarantee of DP-LSVI-UCB^{++}:

Theorem 3.1 (Privacy Guarantee)**.** *DP-LSVI-UCB^{++} (Algorithm 1) satisfies (ϵ, δ')-JDP.*

Proof of Theorem 3.1. For the full proof, refer to Appendix A, particularly Theorem A.1. Note that we use δ' to distinguish between the δ' failure probability of the JDP-mechanism and the δ high probability bounds we get in our regret analysis. At a high level, we first compute the sensitivity of our privatized statistics. With these sensitivities, we simply use a Gaussian mechanism with sufficient noise using Lemma 2.3. Doing this allows us to show that each of our privatized statistics is ρ_0-zCDP so by advanced composition (Lemma 2.2), we can conclde that DP-LSVI-UCB^{++} is ρ-zCDP. Using Lemma 2.1, we can show Algorithm 1 is (ϵ, δ')-DP. Finally, since the actions sent to each user depends on a function constructed with DP and their private data only, we can conclude that DP-LSVI-UCB^{++} is (ϵ, δ')-JDP by the Billboard Lemma (Lemma 2.4). □

Theorem 3.2. *For any linear MDP $\mathcal{M}$, if we set the confidence radii $\widehat{\beta}$, $\check{\beta}$, $\overline{\beta}$ as follows:*

$$\widehat{\beta} = O\left(HL\sqrt{d\tilde{\lambda}_\Lambda} + \sqrt{d^3H^2\log^2\left(\frac{HK^4L^2d}{\delta\tilde{\lambda}_\Lambda}\right)}\right),$$

$$\check{\beta} = O\left(HL\sqrt{d\tilde{\lambda}_\Lambda} + \sqrt{d^3H^2\log^2\left(\frac{HK^4L^2d}{\delta\tilde{\lambda}_\Lambda}\right)}\right),$$

$$\bar{\beta} = O\left(H^2L^2\sqrt{d\tilde{\lambda}_\Lambda} + \sqrt{d^3H^4\log^2\left(\frac{HK^4L^2d}{\delta\tilde{\lambda}_\Lambda}\right)}\right),$$

then with high probability of at least $1 - 7\delta$, the regret of DP-LSVI-UCB^{++} is upper bounded as follows:

$$\textit{Regret}(K) \leq \tilde{O}\left(d\sqrt{H^3K} + \frac{H^{18/4}d^{7/6}K^{1/2}\log(1/\delta')}{\epsilon}\right)$$

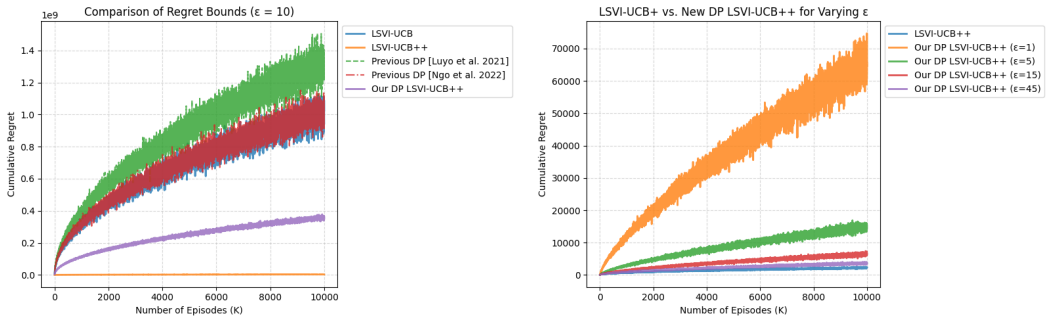
In addition, the number of updates for $\hat{\tilde{Q}}_{k,h}$ and $\check{\tilde{Q}}_{k,h}$ is upper bounded by $O(dH\log(1 + K/d\tilde{\lambda}_\Lambda))$.

Proof of Theorem 3.2. For the full proof, refer to Appendix C, particularly Lemma C.5. At a high level, we replicate the proofs of [He et al. 2023,] with similar function classes for the optimistic,

pessimistic, and squared value functions except using the privatized components. We use these function classes along with standard results for covering numbers to determine the confidence radii (Lemma B.2), prove the upper bound of the variance estimator (Lemma B.3), prove optimism and pessimism, and condition on these to utilize a Bernstein-bound argument to yield a tighter regret bound. Finally, we state some results from [He et al. 2023,] that hold for our analysis and use these to prove the regret bound. □

4 Empirical simulations

We evaluate DP-LSVI-UCB^{++} on a synthetic linear MDP that is described in [Min et al. 2021, Yin et al. 2022, Qiao and Wang 2024,]. In this MDP, we fix the horizon to be $H = 20$. We compare our algorithm compared to LSVI-UCB, their differentially private counterparts proposed by [Luyo et al. 2021, Ngo et al. 2022,], the non-private LSVI-UCB^{++}, and our algorithm DP-LSVI-UCB^{++} in terms of cumulative regret with a fixed privacy budget. We also compare how our regret scales with varying privacy budgets compared to LSVI-UCB^{++}. We ran the simulation 10 times and took the average performance.



(a) Comparison between different algorithms, $H = 20$

(b) Different privacy budgets, $H = 20$

Key Takeaways. From Figure 1a, we can observe the DP-LSVI-UCB^{++} performs better compared to the previous state-of-the-art algorithm devised by [Ngo et al. 2022,] and ofcourse also performs better than [Luyo et al. 2021,]. Additionally, we see that DP-LSVI-UCB^{++} even being a privatized algorithm, performs better than the non-private LSVI-UCB. Looking at Figure 1b, we see that as we increase the privacy budget of DP-LSVI-UCB^{++}, we get closer to LSVI-UCB^{++} and thus with sufficient noise, we can guarantee (ϵ, δ)-JDP that will perform slightly worse than LSVI-UCB^{++}. This is due to the fact that we add Gaussian noise to each count. In particular, we enjoy a better regret bound by using the GOE technique as previous state-of-the-art bounds using a binary-tree mechanism that yields suboptimal regret for the same privacy guarantee. We also enjoy a better bound due to our usage of rare-switching to reduce the amount of noise we added (as adding noise to every statistic would lead to suboptimal regret). This also supports our theoretical regret bound since the cost of privacy appears as lower order terms in the regret bound.

5 Conclusions and future works

In this work, we introduced DP-LSVI-UCB^{++}, a differentially private reinforcement learning algorithm for the linear MDP setting, achieving state-of-the-art regret bounds under joint differential privacy (JDP) constraints. Our approach incorporates advanced techniques such as Bernstein-type martingale concentration inequalities and GOE perturbations, enabling us to improve the utility-privacy tradeoff while maintaining strong theoretical guarantees. The algorithm's regret bound $\tilde{O}\left(\sqrt{H^3K} + H^{19/8}d^{15/8}K^{3/4}/\epsilon\right)$ surpasses prior works and demonstrates that incorporating differential privacy need not lead to substantial utility degradation. Through empirical simulations, we verified that DP-LSVI-UCB^{++} achieves near-optimal performance, often matching or even outperforming non-private baselines. We believe that there are many promising directions to explore from our study. While our work focuses on linear MDPs, it would be valuable to extend these techniques to the low-rank MDP setting. Additionally, our work utilized Gaussian mechanisms and

GOE-based perturbations for privacy guarantees. Exploring alternative mechanisms that adapt noise levels dynamically based on the observed data's sensitivity could lead to improved regret bounds.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, page 2312–2320, Red Hook, NY, USA, 2011. Curran Associates Inc. ISBN 9781618395993.

M Mehdi Afsar, Trafford Crump, and Behrouz Far. Reinforcement learning based recommender systems: A survey. *ACM Computing Surveys*, 55(7):1–38, 2022.

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

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

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. *CoRR*, abs/1605.02065, 2016. URL `http://arxiv.org/abs/1605.02065`.

Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, USA, 2006. ISBN 0521841089.

T.-H. Hubert Chan, Elaine Shi, and Dawn Song. Private and continual release of statistics. *ACM Trans. Inf. Syst. Secur.*, 14(3), November 2011. ISSN 1094-9224. doi: 10.1145/2043621.2043626. URL `https://doi.org/10.1145/2043621.2043626`.

Sayak Ray Chowdhury and Xingyu Zhou. Differentially private regret minimization in episodic markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6375–6383, 2022.

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

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, August 2014. ISSN 1551-305X. doi: 10.1561/0400000042. URL `https://doi.org/10.1561/0400000042`.

Cynthia Dwork and Guy N. Rothblum. Concentrated differential privacy. *CoRR*, abs/1603.01887, 2016. URL `http://arxiv.org/abs/1603.01887`.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Conference on Theory of Cryptography*, TCC'06, page 265–284, Berlin, Heidelberg, 2006a. Springer-Verlag. ISBN 3540327312. doi: 10.1007/11681878_14. URL `https://doi.org/10.1007/11681878_14`.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer, 2006b.

Cynthia Dwork, Moni Naor, Toniann Pitassi, and Guy N. Rothblum. Differential privacy under continual observation. In *Proceedings of the Forty-Second ACM Symposium on Theory of Computing*, STOC '10, page 715–724, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781450300506. doi: 10.1145/1806689.1806787. URL `https://doi.org/10.1145/1806689.1806787`.

John Hartley, Pedro P Sanchez, Fasih Haider, and Sotirios A Tsaftaris. Neural networks memorise personal information from one sample. *Scientific Reports*, 13(1):21366, 2023.

Jiafan He, Heyang Zhao, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal reinforcement learning for linear markov decision processes, 2023. URL https://arxiv.org/abs/2212.06132.

Justin Hsu, Zhiyi Huang, Aaron Roth, Tim Roughgarden, and Zhiwei Steven Wu. Private matchings and allocations. *CoRR*, abs/1311.2828, 2013. URL http://arxiv.org/abs/1311.2828.

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

Michael Kearns, Mallesh M Pai, Ryan Rogers, Aaron Roth, and Jonathan Ullman. Robust mediators in large games. *arXiv preprint arXiv:1512.02698*, 2015.

Abdulrahman Khamaj and Abdulelah M Ali. Adapting user experience with reinforcement learning: Personalizing interfaces based on user behavior analysis in real-time. *Alexandria Engineering Journal*, 95:164–173, 2024.

Mingyang Liu, Xiaotong Shen, and Wei Pan. Deep reinforcement learning for personalized treatment recommendation. *Statistics in medicine*, 41(20):4034–4056, 2022.

Paul Luyo, Evrard Garcelon, Alessandro Lazaric, and Matteo Pirotta. Differentially private exploration in reinforcement learning with linear representation. *arXiv preprint arXiv:2112.01585*, 2021.

Yifei Min, Tianhao Wang, Dongruo Zhou, and Quanquan Gu. Variance-aware off-policy evaluation with linear function approximation. *CoRR*, abs/2106.11960, 2021. URL https://arxiv.org/abs/2106.11960.

Dung Daniel T Ngo, Giuseppe Vietri, and Steven Wu. Improved regret for differentially private exploration in linear mdp. In *International Conference on Machine Learning*, pages 16529–16552. PMLR, 2022.

Dan Qiao and Yu-Xiang Wang. Near-optimal differentially private reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 9914–9940. PMLR, 2023.

Dan Qiao and Yu-Xiang Wang. Offline reinforcement learning with differential privacy. *Advances in Neural Information Processing Systems*, 36, 2024.

Rachel Redberg and Yu-Xiang Wang. Privately publishable per-instance privacy. *CoRR*, abs/2111.02281, 2021. URL https://arxiv.org/abs/2111.02281.

Ahmad EL Sallab, Mohammed Abdou, Etienne Perot, and Senthil Yogamani. Deep reinforcement learning framework for autonomous driving. *arXiv preprint arXiv:1704.02532*, 2017.

Roshan Shariff and Or Sheffet. Differentially private contextual linear bandits. *CoRR*, abs/1810.00068, 2018. URL http://arxiv.org/abs/1810.00068.

Tim van Erven and Peter Harremoës. Rényi divergence and kullback-leibler divergence. *CoRR*, abs/1206.2459, 2012. URL http://arxiv.org/abs/1206.2459.

Giuseppe Vietri, Borja Balle, Akshay Krishnamurthy, and Steven Wu. Private reinforcement learning with pac and regret guarantees. In *International Conference on Machine Learning*, pages 9754–9764. PMLR, 2020.

Tianhao Wang, Dongruo Zhou, and Quanquan Gu. Provably efficient reinforcement learning with linear function approximation under adaptivity constraints. *CoRR*, abs/2101.02195, 2021. URL https://arxiv.org/abs/2101.02195.

Sophia Yazzourh, Nicolas Savy, Philippe Saint-Pierre, and Michael R Kosorok. Medical knowledge integration into reinforcement learning algorithms for dynamic treatment regimes. *arXiv preprint arXiv:2407.00364*, 2024.

Ming Yin, Yaqi Duan, Mengdi Wang, and Yu-Xiang Wang. Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism, 2022. URL https://arxiv.org/abs/2203.05804.

Dongruo Zhou and Quanquan Gu. Computationally efficient horizon-free reinforcement learning for linear mixture mdps, 2022. URL https://arxiv.org/abs/2205.11507.

Xingyu Zhou. Differentially private reinforcement learning with linear function approximation. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 6(1):1–27, 2022.

A Privacy Proofs

First, we will prove the main privacy guarantee of our algorithm

Theorem A.1 (Privacy Guarantee)**.** *DP-LSVI-UCB^{++} (Algorithm 1) satisfies (ϵ, δ)-JDP.*

Proof of Theorem A.1. In order to prove this, we must first determine the l_2 sensitivity of our privatized statistics. Consider two neighboring user sequences $\mathcal{U}$, $\mathcal{U}'$. Let $i \leq k$ be some episode where $s_h^i \neq s_h'^i$ and $a_h^i \neq a_h'^i$ where $\left(s_h^i, a_h^i\right) \in \mathcal{U}$ and $\left(s_h'^i, a_h'^i\right) \in \mathcal{U}'$. Then, $\sum_{i=1}^{k-1} \tilde{\sigma}_{i,h}^{-2} \phi\left(s_h^i, a_h^i\right) \widetilde{V}_{k,h+1}(s_{h+1}^i)$ and $\sum_{i=1}^{k-1} \tilde{\sigma}_{i,h}^{-2} \phi\left(s_h^i, a_h^i\right) \tilde{V}_{k,h+1}(s_{h+1}^i)$ have l_2 sensitivity of $2H$. Likewise, $\sum_{i=1}^{k-1} \tilde{\sigma}_{i,h}^{-2} \phi\left(s_h^i, a_h^i\right) \tilde{V}_{k,h+1}(s_{h+1}^i)^2$ has l_2 sensitivity of $2H^2$. Thus, by using a Gaussian mechanism with noise $\sigma^2 = \frac{2H^2}{\rho_0}$ and $\sigma^2 = \frac{2H^4}{\rho_0}$, respectively, we are guaranteed to have ρ_0-zCDP for each of the first three terms (Lemma 2.3). For the term $\sum_{i=1}^{k-1} \tilde{\sigma}_{i,h}^{-2} \phi(s_i^h, a_i^h) \phi(s_i^h, a_i^h)^\top$, according to Appendix D in [Redberg and Wang 2021,], we have that the per-instance l_2 sensitivity is given as

$$||\Delta_x||_2 = \frac{1}{\sqrt{2}} \text{sup}_{\phi: ||\phi||_2 \leq 1} ||\phi\phi^\top||_F \leq \frac{1}{\sqrt{2}}$$

Thus, by using a Gaussian mechanism with noise $\sigma^2 = \frac{1}{4\rho_0}$, we guarantee that this statistic is ρ_0-zCDP. [2]. Now, we need to track $4KH$ statistics so combing the results of each privatized statistic advanced composition (Lemma 2.2) to conclude that the DP-LSVI-UCB^{++} is ρ-zCDP. Thus, by conversion of zCDP to DP (Lemma 2.1), Algorithm 1 satisfies $\left(\rho + 2\sqrt{\rho \log(1/\delta)}, \delta\right)$-DP. Since the actions sent to each user depends on a function constructed with DP and their private data only, by the Billboard Lemma (Lemma 2.4), we conclude Algorithm 1 is (ϵ, δ)-JDP. □

Now, we will give a high probability bound of the noises we add for privatization. These will be useful for the further analysis we do later on.

Lemma A.1 (Utility Analysis)**.** *Let*

$$L = 4H\sqrt{\frac{dHK}{\rho} \log\left(\frac{10dKH}{\delta}\right)}$$

and

$$\tilde{\lambda}_\Lambda = \sqrt{\frac{8dHK}{\rho}} \left(2 + \left(\frac{\log(5c_1 H/\delta)}{c_2 d}\right)^{\frac{2}{3}}\right)$$

for some universal constants c_1, c_2. Then, with probability atleast $1-\delta$, for all $h, k \in [H] \times [K]$, we have that $||\phi_1||_2 \leq L$, $||\phi_2||_2 \leq L$, and $||\phi_3||_2 \leq HL$. Additionally, we have that K_1 is symmetric and positive definite with $||K_1||_2 \leq \tilde{\lambda}_\Lambda$.

Proof of Lemma A.1. The bounds on ϕ_i hold by simple Gaussian concentration and union bound over all $h, k \in [H] \times [K]$. The bound on K_1 hold from Lemma 19 in [Redberg and Wang 2021,]. □

[2] For those more interested in the details of the GOE DP mechanism, we refer the reader to Appendix D of [Redberg and Wang 2021,]

B Upper Confidence Bound Proofs

We provide this lemma from [Jin et al. 2020,] that we will utilize in our analysis.

Lemma B.1 (Lemma D.1 from [He et al. 2023,], Lemma D.4 from [Jin et al. 2020,] for weighted linear regression)**.** *Let $\{x_k\}_{k=1}^{\infty}$ be a real-valued stochastic process on state space S with corresponding filtration $\{\mathcal{F}_k\}_{k=1}^{\infty}$. Let $\{\phi_k\}_{k=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process, where $\phi_k \in \mathcal{F}_{k-1}$ and $\|\phi_k\|_2 \leq 1$. Let $\{w_k\}_{k=1}^{\infty}$ be a real-valued stochastic process where $w_k \in \mathcal{F}_{k-1}$ and $0 \leq w_k \leq C$. For any $k \geq 0$, define $\Sigma_k = 2\tilde{\lambda}_\Lambda I + \sum_{i=1}^{k} w_i^2 \phi_i \phi_i^\top + K_1$. Then with probability at least $1-\delta$, for all $k \in \mathbb{N}$ and all functions $V \in \mathcal{V}$ with $\max_s |V(x)| \leq H$, we have*

$$\left\| \sum_{i=1}^{k} w_i^2 \phi_i \left\{ V(x_i) - \mathbb{E}[V(x_i)|\mathcal{F}_{i-1}] \right\} \right\|_{\Sigma_k^{-1}}^2 \leq 4C^2H^2 \left[\frac{d}{2} \log\left(1 + \frac{kC^2}{\tilde{\lambda}_\Lambda}\right) + \log\left(\frac{N_\varepsilon}{\delta}\right) \right] + \frac{8k^2C^4\varepsilon^2}{\tilde{\lambda}_\Lambda},$$

where N_ε is the ε-covering number of the function class $\mathcal{V}$ with respect to the distance function dist$(V_1, V_2) = \max_s |V_1(s) - V_2(s)|$.

Proof of Lemma B.1. For any function $V \in \mathcal{V}$, there exists some $\tilde{V}$ in the ε-net such that $\text{dist}\left(V, \tilde{V}\right) \leq \varepsilon$. Using this, the concentration error can be upper bounded as

$$\begin{aligned}\left\| \sum_{i=1}^{k} w_i^2 \phi_i \left\{ V(x_i) - \mathbb{E}[V(x_i)|\mathcal{F}_{i-1}] \right\} \right\|_{\Sigma_k^{-1}}^2 &\leq 2 \left\| \sum_{i=1}^{k} w_i^2 \phi_i \left\{ \tilde{V}(x_i) - \mathbb{E}[\tilde{V}(x_i)|\mathcal{F}_{i-1}] \right\} \right\|_{\Sigma_k^{-1}}^2 \\ &\quad + 2 \left\| \sum_{i=1}^{k} w_i^2 \phi_i \left\{ \Delta_V(x_i) - \mathbb{E}[\Delta_V(x_i)|\mathcal{F}_{i-1}] \right\} \right\|_{\Sigma_k^{-1}}^2\end{aligned}$$

where $\Delta_V = V - \tilde{V}$ and the inequality holds from the fact that $\|a+b\|_\Sigma^2 \leq 2\|a\|_\Sigma^2 + 2\|b\|_\Sigma^2$. For any fixed value function $\tilde{V}$, take $x_i = w_i\phi_i$ and $\eta_i = w_i\tilde{V}(x_i) - w_i\mathbb{E}\left[\tilde{V}(x_i)\right]$. Notice that

$$\begin{aligned}&\|x_i\|_2 \leq C \\ &\mathbb{E}\left[\eta_i \mid \mathcal{F}_i\right] = 0,\ |\eta_i| \leq HC\end{aligned}$$

Then, by Lemma G.6 and taking a union-bound over the ε-net $\mathcal{N}_\varepsilon$, we get the first term being upper bounded as

$$2 \left\| \sum_{i=1}^{k} w_i^2 \phi_i \left\{ \tilde{V}(x_i) - \mathbb{E}[\tilde{V}(x_i)|\mathcal{F}_{i-1}] \right\} \right\|_{\Sigma_k^{-1}}^2 \leq 4H^2C^2 \left[\frac{d}{2} \log\left(1 + KC^2/\tilde{\lambda}_\Lambda\right) + \log\frac{N_\varepsilon}{\delta} \right]$$

The second term can be upper bounded as

$$\begin{aligned}2 \left\| \sum_{i=1}^{k} w_i^2 \phi_i \left\{ \Delta_V(x_i) - \mathbb{E}[\Delta_V(x_i)|\mathcal{F}_{i-1}] \right\} \right\|_{\Sigma_k^{-1}}^2 &\leq 2k \sum_{i=1}^{k} \left\| w_i^2 \phi_i \left\{ \Delta_V(x_i) - \mathbb{E}[\Delta_V(x_i)|\mathcal{F}_{i-1}] \right\} \right\|_{\Sigma_k^{-1}}^2 \\ &\leq 8k^2C^4\varepsilon^2/\tilde{\lambda}_\Lambda\end{aligned}$$

where the first inequality holds from Cauchy-Schwartz and the last inequality holds from $|\Delta_V| \leq \varepsilon$, $w_i^2 \leq C^2$, and $\Sigma_k \succeq \tilde{\lambda}_\Lambda$. Thus, putting these together, we get the claim.

□

Now, we are ready to begin to derive our confidence radii. This is a Hoeffding-type upper bound for the estimation error.

Lemma B.2. *Define $\mathcal{E}$ as the event that the following inequalities hold for all $s, a, k, h \in \mathcal{S} \times \mathcal{A} \times [K] \times [H]$:*

$$\left|\tilde{\widehat{w}}_{k,h}^{\top}\phi(s,a) - [\mathbb{P}_h\tilde{\widehat{V}}_{k,h+1}](s,a)\right| \leq \widehat{\beta}\sqrt{\phi(s,a)^{\top}\tilde{\Lambda}_{k,h}^{-1}\phi(s,a)},$$

$$\left|\tilde{\bar{w}}_{k,h}^{\top}\phi(s,a) - [\mathbb{P}_h\tilde{\widehat{V}}^2_{k,h+1}](s,a)\right| \leq \bar{\beta}\sqrt{\phi(s,a)^{\top}\tilde{\Lambda}_{k,h}^{-1}\phi(s,a)},$$

$$\left|\tilde{\breve{w}}_{k,h}^{\top}\phi(s,a) - [\mathbb{P}_h\tilde{\breve{V}}_{k,h+1}](s,a)\right| \leq \breve{\beta}\sqrt{\phi(s,a)^{\top}\tilde{\Lambda}_{k,h}^{-1}\phi(s,a)},$$

where

$$\widehat{\beta} = \breve{\beta} = O\left(HL\sqrt{d\tilde{\lambda}_{\Lambda}} + \sqrt{d^3H^2\log^2\left(\frac{HK^4L^2d}{\delta\tilde{\lambda}_{\Lambda}}\right)}\right),$$

and

$$\bar{\beta} = O\left(H^2L^2\sqrt{d\tilde{\lambda}_{\Lambda}} + \sqrt{d^3H^4\log^2\left(\frac{HK^4L^2d}{\delta\tilde{\lambda}_{\Lambda}}\right)}\right).$$

The event $\mathcal{E}$ holds with probability at least $1 - 7\delta$.

Proof of Lemma B.2. For any fixed stage $h \in [H]$ and the optimistic private value function $\tilde{\widehat{V}}_{k,h+1}$, by Lemma G.1, there exists a vector $w_{k,h+1}$ such that $\mathbb{P}_h\tilde{\widehat{V}}_{k,h+1}(s,a)$ can be represented as $w_{k,h}^{\top}\phi(s,a)$ with $\|w_{k,h}\|_2 \leq H\sqrt{d}$. Then, we can decompose the estimation error $\left\|\tilde{\widehat{w}}_{k,h} - w_{k,h}\right\|_{\tilde{\Lambda}_{h,k}}$ as

$$\begin{aligned}
&\left\|\tilde{\Lambda}_{k,h}^{-1}\left[\sum_{i=1}^{k-1}\tilde{\widehat{\sigma}}_{i,h}^{-2}\phi\left(s_h^i,a_h^i\right)\tilde{\widehat{V}}_{k,h+1}(s_{h+1}^i) + \phi_1\right] - \tilde{\Lambda}_{k,h}^{-1}\left[2\tilde{\lambda}_{\Lambda}I_d + \sum_{i=1}^{k-1}\tilde{\widehat{\sigma}}_{i,h}^{-2}\phi\left(s_h^i,a_h^i\right)\phi\left(s_h^i,a_h^i\right)^{\top} + K_1\right]w_{k,h}\right\|_{\tilde{\Lambda}_{h,k}} \\
&= \left\|\tilde{\Lambda}_{k,h}^{-1}\sum_{i=1}^{k-1}\tilde{\widehat{\sigma}}_{i,h}^{-2}\phi\left(s_h^i,a_h^i\right)\left(\tilde{\widehat{V}}_{k,h+1}(s_{h+1}^i) - \mathbb{P}_h\tilde{\widehat{V}}_{k,h+1}(s_h^i,a_h^i)\right) + \tilde{\Lambda}_{k,h}^{-1}\phi_1 + \tilde{\Lambda}_{k,h}^{-1}K_1w_{k,h} - 2\tilde{\lambda}_{\Lambda}\tilde{\Lambda}_{k,h}^{-1}w_{k,h}\right\|_{\tilde{\Lambda}_{h,k}} \\
&\leq \left\|\tilde{\Lambda}_{k,h}^{-1}\phi_1\right\|_{\tilde{\Lambda}_{h,k}} + \left\|\tilde{\Lambda}_{k,h}^{-1}w_{k,h}K_1\right\|_{\tilde{\Lambda}_{h,k}} + \left\|2\tilde{\lambda}_{\Lambda}\tilde{\Lambda}_{k,h}^{-1}w_{k,h}\right\|_{\tilde{\Lambda}_{h,k}} + \\
&\left\|\tilde{\Lambda}_{k,h}^{-1}\sum_{i=1}^{k-1}\tilde{\widehat{\sigma}}_{i,h}^{-2}\phi\left(s_h^i,a_h^i\right)\left(\tilde{\widehat{V}}_{k,h+1}(s_{h+1}^i) - \mathbb{P}_h\tilde{\widehat{V}}_{k,h+1}(s_h^i,a_h^i)\right)\right\|_{\tilde{\Lambda}_{h,k}}
\end{aligned}$$

where the first inequality holds from $\|a+b\|_{\Sigma} \leq \|a\|_{\Sigma} + \|b\|_{\Sigma}$. For the first term, we know that by construction, $\tilde{\Lambda}_{k,h}^{-1} \preceq 1/\tilde{\lambda}_{\Lambda}$. Additionally, by utility (Lemma A.1), we have that $\|\phi_1\|_2 \leq L$. Putting these together, we get

$$\left\|\tilde{\Lambda}_{k,h}^{-1}\phi_1\right\|_{\tilde{\Lambda}_{h,k}} \leq L\sqrt{\frac{1}{\tilde{\lambda}_{\Lambda}}} \leq HL\sqrt{d\tilde{\lambda}_{\Lambda}}$$

For the second term, we have that $\|w_{k,h}\|_2 \leq H\sqrt{d}$. Again, by utility, we have that $\|K_1\|_2 \leq \tilde{\lambda}_{\Lambda}$. Thus, we get

$$\left\|\tilde{\Lambda}_{k,h}^{-1}w_{k,h}K_1\right\|_{\tilde{\Lambda}_{h,k}} \leq H\sqrt{d\tilde{\lambda}_{\Lambda}} \leq HL\sqrt{d\tilde{\lambda}_{\Lambda}}$$

For the third term, using the facts we have described above, we get

$$\left\|2\tilde{\lambda}_{\Lambda}\tilde{\Lambda}_{k,h}^{-1}w_{k,h}\right\|_{\tilde{\Lambda}_{h,k}} \leq 2H\sqrt{d\tilde{\lambda}_{\Lambda}} \leq 2HL\sqrt{d\tilde{\lambda}_{\Lambda}}$$

Lastly, for the last term, we apply Lemma B.1 with the following optimistic value function class $\widehat{\mathcal{V}}_h$ and $\varepsilon = H\sqrt{d\tilde{\lambda}_{\Lambda}}/K$, then for any fixed $h \in [H]$, with probability atleast $1 - \delta/H$, for all episodes

$k \in [K]$, we have

$$\begin{aligned}
&\left\| \tilde{\Lambda}_{k,h}^{-1} \sum_{i=1}^{k-1} \tilde{\bar{\sigma}}_{i,h}^{-2} \phi\left(s_h^i, a_h^i\right) \left(\tilde{\widehat{V}}_{k,h+1}(s_{h+1}^i) - \mathbb{P}_h \tilde{\widehat{V}}_{k,h+1}(s_h^i, a_h^i) \right) \right\|_{\tilde{\Lambda}_{h,k}} \\
&\leq \sqrt{4C^2H^2 \left[\frac{d}{2} \log\left(1 + \frac{kC^2}{\tilde{\lambda}_\Lambda}\right) + \log\left(\frac{HN_\varepsilon}{\delta}\right) \right] + \frac{8k^2C^4\varepsilon^2}{\tilde{\lambda}_\Lambda}} \\
&\leq \sqrt{4H \left[\frac{d}{2} \log\left(1 + \frac{k}{\tilde{\lambda}_\Lambda H}\right) + \log\left(\frac{HN_\varepsilon}{\delta}\right) \right] + \frac{8k^2\varepsilon^2}{\tilde{\lambda}_\Lambda H^2}} \\
&\leq \sqrt{4H \left[\frac{d}{2} \log\left(1 + \frac{k}{\tilde{\lambda}_\Lambda H}\right) + \log\left(\frac{HN_\varepsilon}{\delta}\right) \right] + 8} \\
&= O\left(\sqrt{d^3H^2 \log^2\left(\frac{HK^4L^2d}{\delta\tilde{\lambda}_\Lambda}\right)} \right)
\end{aligned}$$

where the first inequality holds due to Lemma B.1, the second inequality holds since $\tilde{\bar{\sigma}}_{i,h}^{-2} \leq 1/\sqrt{H}$, and the last inequality holds due to Lemma F.2 and $\varepsilon = H\sqrt{d\tilde{\lambda}_\Lambda}/K$. Putting everything together, we get

$$\left\| \tilde{\widehat{w}}_{k,h} - w_{k,h} \right\|_{\tilde{\Lambda}_{h,k}} \leq O\left(HL\sqrt{d\tilde{\lambda}_\Lambda} + \sqrt{d^3H^2 \log^2\left(\frac{HK^4L^2d}{\delta\tilde{\lambda}_\Lambda}\right)} \right) = \widehat{\beta}$$

Thus, using this, we can say

$$\begin{aligned}
\left| \tilde{\widehat{w}}_{k,h}^\top \phi(s,a) - [\mathbb{P}_h \tilde{\widehat{V}}_{k,h+1}](s,a) \right| &= \left| \tilde{\widehat{w}}_{k,h}^\top \phi(s,a) - w_{k,h}^\top \phi(s,a) \right| \\
&\leq \left\| \tilde{\widehat{w}}_{k,h} - w_{k,h} \right\|_{\tilde{\Lambda}_{h,k}} \|\phi(s,a)\|_{\tilde{\Lambda}_{h,k}} \\
&\leq \widehat{\beta} \sqrt{\phi(s,a)^\top \tilde{\Lambda}_{k,h}^{-1} \phi(s,a)}
\end{aligned}$$

where the first inequality holds due to Cauchy-Schwartz inequality. Replacing the value function class by the pessimistic value function class $\check{\mathcal{V}}$ or the squared value function class $\widehat{\mathcal{V}}^2$ and using the same exact proof as above, we can derive the other upper estimation errors

$$\left| \tilde{\widehat{w}}_{k,h}^\top \phi(s,a) - [\mathbb{P}_h \tilde{\widehat{V}}_{k,h+1}^2](s,a) \right| \leq \bar{\beta} \sqrt{\phi(s,a)^\top \tilde{\Lambda}_{k,h}^{-1} \phi(s,a)},$$

$$\left| \tilde{\check{w}}_{k,h}^\top \phi(s,a) - [\mathbb{P}_h \tilde{\check{V}}_{k,h+1}](s,a) \right| \leq \check{\beta} \sqrt{\phi(s,a)^\top \tilde{\Lambda}_{k,h}^{-1} \phi(s,a)},$$

where

$$\check{\beta} = O\left(HL\sqrt{d\tilde{\lambda}_\Lambda} + \sqrt{d^3H^2 \log^2\left(\frac{HK^4L^2d}{\delta\tilde{\lambda}_\Lambda}\right)} \right),$$

and

$$\bar{\beta} = O\left(H^2L^2\sqrt{d\tilde{\lambda}_\Lambda} + \sqrt{d^3H^4 \log^2\left(\frac{HK^4L^2d}{\delta\tilde{\lambda}_\Lambda}\right)} \right).$$

□

Now, we provide a bound on the variance estimator.

Lemma B.3. *Let $\widetilde{\mathcal{E}}_h$ be the event such that for all episodes $k \in [K]$, stages $h \leq h' \leq H$, and state-action pairs $(s,a) \in \mathcal{S} \times \mathcal{A}$, the weight vector $\widehat{w}_{k,h}$ satisfies*

$$\left| \tilde{\widehat{w}}_{k,h'}^\top \phi(s,a) - [\mathbb{P}_h \tilde{\widehat{V}}_{k,h'+1}](s,a) \right| \leq \beta \sqrt{\phi(s,a)^\top \tilde{\Lambda}_{k,h'}^{-1} \phi(s,a)} \tag{B.1}$$

where

$$\beta = O\left(HL\sqrt{d\tilde{\lambda}_\Lambda} + \sqrt{d\log^2\left(1+\left(\frac{HK^4L^2d}{\delta\tilde{\lambda}_\Lambda}\right)\right)}\right).$$

On the event $\mathcal{E}$ and $\widetilde{\mathcal{E}}_{h+1}$, for each episode $k \in [K]$ and stage h, the estimated variance satisfies:

$$\left|[\overline{\mathbb{V}}_h\tilde{\widehat{V}}_{k,h+1}](s_k^h,a_k^h) - [\mathbb{V}_h\tilde{\widehat{V}}_{k,h+1}](s_k^h,a_k^h)\right| \le E_{k,h},$$

and

$$\left|[\overline{\mathbb{V}}_h\tilde{\widehat{V}}_{k,h+1}](s_h^k,a_h^k) - [\mathbb{V}_hV_h^*](s_h^k,a_h^k)\right| \le E_{k,h} + D_{k,h}.$$

where

$$E_{k,h} = \min\left\{\overline{\beta}_k\left\|\tilde{\Lambda}_{k,h}^{-1/2}\phi(s_h^k,a_h^k)\right\|_2, H^2\right\} + \min\left\{2H\widehat{\beta}_k\left\|\tilde{\Lambda}_{k,h}^{-1/2}\phi(s_h^k,a_h^k)\right\|_2, H^2\right\}$$

and

$$D_{k,h} = \min\left\{4d^3H^2\left(\tilde{\widehat{w}}_{k,h}^\top\phi(s,a) - \tilde{\breve{w}}_{k,h}^\top\phi(s,a) + 2\widehat{\beta}_k\sqrt{\phi(s,a)^\top\tilde{\Lambda}_{k,h}^{-1}\phi(s,a)}\right), d^3H^3\right\}$$

Proof of Lemma B.3. We will first use Lemma B.2

$$\begin{aligned}
&\left|[\overline{\mathbb{V}}_h\tilde{\widehat{V}}_{k,h+1}](s_k^h,a_k^h) - [\mathbb{V}_h\tilde{\widehat{V}}_{k,h+1}](s_k^h,a_k^h)\right| \\
&= \left|[\tilde{\overline{w}}_{k,h}\phi(s_h^k,a_h^k)]_{[0,H^2]} - [\tilde{\widehat{w}}_{k,h}\phi(s_h^k,a_h^k)]_{[0,H]}^2 - [\mathbb{P}_h\tilde{\widehat{V}}_{k,h+1}^2](s_h^k,a_h^k) - \left([\mathbb{P}_h\tilde{\widehat{V}}_{k,h+1}]](s_h^k,a_h^k)\right)^2\right| \\
&\le \left|[\tilde{\overline{w}}_{k,h}\phi(s_h^k,a_h^k)]_{[0,H^2]} - [\mathbb{P}_h\tilde{\widehat{V}}_{k,h+1}^2](s_h^k,a_h^k)\right| + \left|[\tilde{\widehat{w}}_{k,h}\phi(s_h^k,a_h^k)]_{[0,H]}^2 - \left([\mathbb{P}_h\tilde{\widehat{V}}_{k,h+1}]_{k,h+1}](s_h^k,a_h^k)\right)^2\right| \\
&= \left|[\tilde{\overline{w}}_{k,h}\phi(s_h^k,a_h^k)]_{[0,H^2]} - [\mathbb{P}_h\tilde{\widehat{V}}_{k,h+1}]_{k,h+1}^2](s_h^k,a_h^k)\right| \\
&+ \left|[\tilde{\widehat{w}}_{k,h}\phi(s_h^k,a_h^k)]_{[0,H]} + [\mathbb{P}_h\tilde{\widehat{V}}_{k,h+1}]_{k,h+1}](s_h^k,a_h^k)\right|\left|[\tilde{\widehat{w}}_{k,h}\phi(s_h^k,a_h^k)]_{[0,H]} - [\mathbb{P}_h\tilde{\widehat{V}}_{k,h+1}]_{k,h+1}](s_h^k,a_h^k)\right| \\
&\le \min\left\{\overline{\beta}_k\left\|\tilde{\Lambda}_{k,h}^{-1/2}\phi(s_h^k,a_h^k)\right\|_2, H^2\right\} + \min\left\{2H\widehat{\beta}_k\left\|\tilde{\Lambda}_{k,h}^{-1/2}\phi(s_h^k,a_h^k)\right\|_2, H^2\right\} \\
&= E_{k,h}
\end{aligned}$$

where the first inequality holds from the triangle inequality and the second inequality holds from conditioning on $\mathcal{E}$ and the fact that $0 \le [\tilde{\widehat{w}}_{k,h}\phi_{s_h^k,s_h^k}]_{[0,H]} + [\mathbb{P}_hV_{k,h+1}](s_h^k,a_h^k) \le 2H$. Now,

$$\begin{aligned}
&\left|[\mathbb{V}_h\tilde{\widehat{V}}_{k,h+1}](s_h^k,a_h^k) - [\mathbb{V}_hV_h^*](s_h^k,a_h^k)\right| \\
&= \left|[\mathbb{P}_h\tilde{\widehat{V}}_{k,h+1}^2](s_h^k,a_h^k) - \left([\mathbb{P}_h\tilde{\widehat{V}}_{k,h+1}](s_h^k,a_h^k)\right)^2 - [\mathbb{P}_hV_{h+1}^{*2}](s_h^k,a_h^k) + \left([\mathbb{P}_hV_{h+1}^*](s_h^k,a_h^k)\right)^2\right| \\
&\le \left|[\mathbb{P}_h\left(\tilde{\widehat{V}}_{k,h+1} - V_{h+1}^*\right)\left(\tilde{\widehat{V}}_{k,h+1} + V_{h+1}^*\right)](s_h^k,a_h^k)\right| \\
&+ \left|\left([\mathbb{P}_h\tilde{\widehat{V}}_{h+1}](s_h^k,a_h^k) - [\mathbb{P}_hV_{h+1}^*](s_h^k,a_h^k)\right)\left([\mathbb{P}_h\tilde{\widehat{V}}_{k,h+1}](s_h^k,a_h^k) + [\mathbb{P}_hV_{h+1}^*](s_h^k,a_h^k)\right)\right| \\
&\le 4H\left([\mathbb{P}_h\tilde{\widehat{V}}_{k,h+1}](s_h^k,a_h^k) - [\mathbb{P}_hV_{h+1}^*](s_h^k,a_h^k)\right)
\end{aligned}$$

where the first inequality holds from triangle inequality and the second inequality holds due to Lemma B.5 and the fact that $0 \le V_{h+1}^*(s' \le V_{k,h+1}(s') \le H$. Now, if we condition on $\mathcal{E}$ and $\tilde{\mathcal{E}}$, we get

$$\begin{aligned}
&\left([\mathbb{P}_h\tilde{\widehat{V}}_{k,h+1}](s_h^k,a_h^k) - [\mathbb{P}_hV_{h+1}^*](s_h^k,a_h^k)\right) \\
&\le \left([\mathbb{P}_h\tilde{\widehat{V}}_{k,h+1}](s_h^k,a_h^k) - [\mathbb{P}_h\tilde{\breve{V}}_{k,h+1}](s_h^k,a_h^k)\right) \\
&\le \tilde{\widehat{w}}_{k,h}^\top\phi(s,a) + \widehat{\beta}\sqrt{\phi(s,a)^\top\tilde{\Lambda}_{k,h}^{-1}\phi(s,a)} - \tilde{\breve{w}}_{k,h}^\top\phi(s,a) + \breve{\beta}\sqrt{\phi(s,a)^\top\tilde{\Lambda}_{k,h}^{-1}\phi(s,a)}
\end{aligned}$$

where the first inequality holds due to Lemma B.5 and the last inequality holds by Lemma B.2. Combining results, we get

$$\begin{aligned}
&\left|[\overline{\mathbb{V}}_h\tilde{\hat{V}}_{k,h+1}](s_h^k,a_h^k)-[\mathbb{V}_hV_{h+1}^*]\right| \\
&\leq \left|[\overline{\mathbb{V}}_h\tilde{\hat{V}}_{k,h+1}](s_h^k,a_h^k)-[\mathbb{V}_h\tilde{\hat{V}}_{k,h+1}^*](s_h^k,a_h^k)\right|+\left|[\mathbb{V}_h\tilde{\hat{V}}_{k,h+1}](s_h^k,a_h^k)-[\mathbb{V}_hV_{h+1}^*](s_h^k,a_h^k)\right| \\
&\leq E_{k,h}+\min\left\{4H\left(\tilde{\hat{w}}_{k,h}^\top\phi(s,a)+\hat{\beta}\sqrt{\phi(s,a)^\top\tilde{\Lambda}_{k,h}^{-1}\phi(s,a)}-\tilde{\check{w}}_{k,h}^\top\phi(s,a)+\check{\beta}\sqrt{\phi(s,a)^\top\tilde{\Lambda}_{k,h}^{-1}\phi(s,a)}\right),H^2\right\}
\end{aligned}$$

□

We also have another upper bound on the variance estimator.

Lemma B.4. *On the event $\mathcal{E}$ and $\tilde{\mathcal{E}}_{h+1}$, for any episode k and $i>k$, we have*

$$[\mathbb{V}_h(\tilde{\hat{V}}_{i,h+1}-V_{h+1}^*)](s_h^k,a_h^k)\leq D_{k,h}/d^3H.$$

Proof of Lemma B.4. On the event $\mathcal{E}$ and $\tilde{\mathcal{E}}_{h+1}$, we have

$$\begin{aligned}
[\mathbb{V}_h(\tilde{\hat{V}}_{i,h+1}-\mathbb{V}_{h+1}^*)](s_h^k,a_h^k)&\leq[\mathbb{P}_h(\tilde{\hat{V}}_{i,h+1}-V_{h+1}^*)^2](s_h^k,a_h^k) \\
&\leq 2H[\mathbb{P}_h(\tilde{\hat{V}}_{i,h+1}-V_{h+1}^*)](s_h^k,a_h^k) \\
&\leq 2H\left([\mathbb{P}_h(\tilde{\hat{V}}_{i,h+1}](s_h^k,a_h^k)-[\mathbb{P}_h(\tilde{\check{V}}_{i,h+1}](s_h^k,a_h^k)\right) \\
&\leq 2H\left([\mathbb{P}_h(\tilde{\hat{V}}_{k,h+1}](s_h^k,a_h^k)-[\mathbb{P}_h(\tilde{\check{V}}_{i,h+1}](s_h^k,a_h^k)\right) \\
&\leq 2H\left(\tilde{\hat{w}}_{k,h}^\top\phi(s,a)+\hat{\beta}\sqrt{\phi(s,a)^\top\tilde{\Lambda}_{k,h}^{-1}\phi(s,a)}-\tilde{\check{w}}_{k,h}^\top\phi(s,a)+\check{\beta}\sqrt{\phi(s,a)^\top\tilde{\Lambda}_{k,h}^{-1}\phi(s,a)}\right)
\end{aligned}$$

where the first inequality holds due to $\mathrm{Var}(x)\leq\mathbb{E}[x^2]$, the second and third inequalities hold due to Lemma B.5 with the fact that $0\leq\tilde{\hat{V}}_{i,h+1}(s'),V_{h+1}^*(s')\leq H$, the fourth inequality holds because $V_{k,h+1}\geq V_{i,h+1}$ from the update rule in Algorithm 1, and the fifth inequality holds due to Lemma B.2. On the other hand, since the value functions satisfy $0\leq\tilde{\hat{V}}_{i,h+1}(s'),V_{h+1}^*(s')\leq H$, we have

$$[V_h(V_{i,h+1}-V_{h+1}^*)](s_h^k,a_h^k)\leq\frac{d^3H^3}{d^3H}=H^2.$$

□

Here, we prove the optimism and pessimism of our privatized value function which we will use for the regret analysis.

Lemma B.5 (Privatized Optimism and Pessimism)**.** *On the event $\mathcal{E}$ and $\tilde{\mathcal{E}}_h$, for all episodes $k\in[K]$ and stages $h\leq h'\leq H$, we have*

$$\tilde{\hat{Q}}_{k,h}(s,a)\geq Q_h^*(s,a)\geq\tilde{\check{Q}}_{k,h}(s,a).$$

In addition, we have

$$\tilde{\hat{V}}_{k,h}(s)\geq V_h^*(s)\geq\tilde{\check{V}}_{k,h}(s).$$

Proof of Lemma B.5. As we would usually do, we will prove optimism and pessimism using induction. First, consider the base case $H+1$. For all states $s\in S$ and actions $a\in A$, we have

$$\tilde{\hat{Q}}_{k,H+1}(s,a)=Q_h^*(s,a)=\tilde{\check{Q}}_{k,h}(s,a)=0\quad\text{and}\quad\tilde{\hat{V}}_{k,h}(s)\geq V_h^*(s)\geq\tilde{\check{V}}_{k,h}(s)=0.$$

Thus, we have shown the base case. Now, consider stage $h+1$. Since the event $\tilde{\mathcal{E}}_h$ directly implies the event $\tilde{\mathcal{E}}_{h+1}$, according to the induction hypothesis, we have

$$\widehat{\tilde{V}}_{k,h+1}(s) \geq V^*_{h+1}(s) \geq \check{\tilde{V}}_{k,h+1}(s).$$

Thus, for all episodes $k \in [K]$, we have

$$r_h(s,a) + \widehat{\tilde{w}}_k^\top \phi(s,a) + \widehat{\beta}\sqrt{\phi(s,a)^\top \tilde{\Lambda}_{k,h}^{-1}\phi(s,a)} - Q^*_h(s,a) \geq [\mathbb{P}_h(\widehat{\tilde{V}}_{k,h+1} - V^*_{h+1})](s,a) \geq 0,$$

where the first inequality holds by conditioning on event $\tilde{\mathcal{E}}_h$. Additionally, we have

$$Q^*_h(s,a) \leq \min\left\{\min_{1\leq i\leq k}\left(r_h(s,a) + \widehat{\tilde{w}}_i^\top \phi(s,a) + \widehat{\beta}\sqrt{\phi(s,a)^\top \tilde{\Lambda}_{i,h}^{-1}\phi(s,a)}\right), H\right\} \leq \widehat{\tilde{Q}}_{k,h}(s,a).$$

With a similar argument, for the pessimistic action-value function $\check{\tilde{Q}}_{k,h}(s,a)$, we have

$$r_h(s,a) + \check{\tilde{w}}_k^\top \phi(s,a) - \check{\beta}\sqrt{\phi(s,a)^\top \tilde{\Lambda}_{k,h}^{-1}\phi(s,a)} - Q^*_h(s,a) \leq [\mathbb{P}_h(\check{\tilde{V}}_{k,h+1} - V^*_{h+1})](s,a) \leq 0.$$

Since the optimal value function is lower bounded by $Q^*_h(s,a) \geq 0$, the result further implies that

$$Q^*_h(s,a) \geq \max\left\{\max_{1\leq i\leq k}\left(r_h(s,a) + \check{\tilde{w}}_{\text{last},h}^\top \phi(s,a) + \check{\beta}\sqrt{\phi(s,a)^\top \tilde{\Lambda}_{\text{last},h}^{-1}\phi(s,a)}\right), 0\right\} \geq \check{\tilde{Q}}_{k,h}(s,a).$$

In addition, we have

$$\widehat{\tilde{V}}_{k,h}(s) = \max_a \widehat{\tilde{Q}}_{i,h}(s,a) \geq \min_{1\leq i\leq k}\max_a Q^*_h(s,a) = V^*_h(s),$$

$$\check{\tilde{V}}_{k,h}(s) = \max_a \widehat{\tilde{Q}}_{i,h}(s,a) \leq \max_{1\leq i\leq k}\max_a Q^*_h(s,a) = V^*_h(s),$$

□

Now, we will also provide a Bernstein-type upper bound on the estimation error using what have proven so far. This is much sharper than Lemma B.2.

Lemma B.6. *Define $\widetilde{\mathcal{E}} = \widetilde{\mathcal{E}}_1$ as the event such that B.1 holds for all stages $h \in [H]$. On the events $\mathcal{E}$, event $\tilde{\mathcal{E}}$ holds with probability at least $1-\delta$*

Proof. For any fixed stage $h \in [H]$ and the optimistic private value function $\widehat{\tilde{V}}_{k,h+1}$, by Lemma G.1, there exists a vector $w_{k,h+1}$ such that $\mathbb{P}_h\widehat{\tilde{V}}_{k,h+1}(s,a)$ can be represented as $w_{k,h}^\top\phi(s,a)$ with $\|w_{k,h}\|_2 \leq H\sqrt{d}$. Then, we can decompose the estimation error $\left\|\widehat{\tilde{w}}_{k,h} - w_{k,h}\right\|_{\tilde{\Lambda}_{h,k}}$ as

$$\begin{aligned}
&\left\|\tilde{\Lambda}_{k,h}^{-1}\left[\sum_{i=1}^{k-1}\tilde{\tilde{\sigma}}_{i,h}^{-2}\phi\left(s_h^i,a_h^i\right)\widehat{\tilde{V}}_{k,h+1}(s_{h+1}^i) + \phi_1\right] - \tilde{\Lambda}_{k,h}^{-1}\left[2\tilde{\lambda}_\Lambda I_d + \sum_{i=1}^{k-1}\tilde{\tilde{\sigma}}_{i,h}^{-2}\phi\left(s_h^i,a_h^i\right)\phi\left(s_h^i,a_h^i\right)^\top + K_1\right]w_{k,h}\right\|_{\tilde{\Lambda}_{h,k}} \\
&= \left\|\tilde{\Lambda}_{k,h}^{-1}\sum_{i=1}^{k-1}\tilde{\tilde{\sigma}}_{i,h}^{-2}\phi\left(s_h^i,a_h^i\right)\left(\widehat{\tilde{V}}_{k,h+1}(s_{h+1}^i) - \mathbb{P}_h\widehat{\tilde{V}}_{k,h+1}(s_h^i,a_h^i)\right) + \tilde{\Lambda}_{k,h}^{-1}\phi_1 + \tilde{\Lambda}_{h,k}^{-1}w_{k,h}K_1 - 2\tilde{\lambda}_\Lambda\tilde{\Lambda}_{k,h}^{-1}w_{k,h}\right\|_{\tilde{\Lambda}_{h,k}} \\
&\leq \left\|\tilde{\Lambda}_{k,h}^{-1}\phi_1\right\|_{\tilde{\Lambda}_{h,k}} + \left\|\tilde{\Lambda}_{h,k}^{-1}w_{k,h}K_1\right\|_{\tilde{\Lambda}_{h,k}} + \left\|2\tilde{\lambda}_\Lambda\tilde{\Lambda}_{k,h}^{-1}w_{k,h}\right\|_{\tilde{\Lambda}_{h,k}} + \\
&\left\|\tilde{\Lambda}_{k,h}^{-1}\sum_{i=1}^{k-1}\tilde{\tilde{\sigma}}_{i,h}^{-2}\phi\left(s_h^i,a_h^i\right)\left(\widehat{\tilde{V}}_{k,h+1}(s_{h+1}^i) - \mathbb{P}_h\widehat{\tilde{V}}_{k,h+1}(s_h^i,a_h^i)\right)\right\|_{\tilde{\Lambda}_{h,k}}
\end{aligned}$$

where the first inequality holds from $\|a+b\|_\Sigma \le \|a\|_\Sigma + \|b\|_\Sigma$. For the first term, we know that by construction, $\tilde{\Lambda}_{k,h}^{-1} \preceq 1/\tilde{\lambda}_\Lambda$. Additionally, by utility (Lemma A.1), we have that $\|\phi_1\|_2 \le L$. Putting these together, we get

$$\left\|\tilde{\Lambda}_{k,h}^{-1}\phi_1\right\|_{\tilde{\Lambda}_{h,k}} \le L\sqrt{\frac{1}{\tilde{\lambda}_\Lambda}} \le HL\sqrt{d\tilde{\lambda}_\Lambda}$$

For the second term, we have that $\|w_{k,h}\|_2 \le H\sqrt{d}$. Again, by utility, we have that $\|K_1\|_2 \le \tilde{\lambda}_\Lambda$. Thus, we get

$$\left\|\tilde{\Lambda}_{h,k}^{-1}w_{k,h}K_1\right\|_{\tilde{\Lambda}_{h,k}} \le H\sqrt{d\tilde{\lambda}_\Lambda} \le HL\sqrt{d\tilde{\lambda}_\Lambda}$$

For the third term, using the facts we have described above, we get

$$\left\|2\tilde{\lambda}_\Lambda\tilde{\Lambda}_{k,h}^{-1}w_{k,h}\right\|_{\tilde{\Lambda}_{h,k}} \le 2H\sqrt{d\tilde{\lambda}_\Lambda} \le 2HL\sqrt{d\tilde{\lambda}_\Lambda}$$

For the last term,

$$\begin{aligned}
&\left\|\tilde{\Lambda}_{k,h}^{-1}\sum_{i=1}^{k-1}\tilde{\tilde{\sigma}}_{i,h}^{-2}\phi\left(s_h^i,a_h^i\right)\left(\hat{\tilde{V}}_{k,h+1}(s_{h+1}^i)-\mathbb{P}_h\hat{\tilde{V}}_{k,h+1}(s_h^i,a_h^i)\right)\right\|_{\tilde{\Lambda}_{h,k}}\\
&=\left\|\sum_{i=1}^{k-1}\tilde{\tilde{\sigma}}_{i,h}^{-2}\phi\left(s_h^i,a_h^i\right)\left(\hat{\tilde{V}}_{k,h+1}(s_{h+1}^i)-\mathbb{P}_h\hat{\tilde{V}}_{k,h+1}(s_h^i,a_h^i)\right)\right\|_{\tilde{\Lambda}_{h,k}^{-1}}\\
&\le\left\|\sum_{i=1}^{k-1}\tilde{\tilde{\sigma}}_{i,h}^{-2}\phi\left(s_h^i,a_h^i\right)\left(V_{h+1}^*(s_{h+1}^i)-\mathbb{P}_hV_{h+1}^*(s_h^i,a_h^i)\right)\right\|_{\tilde{\Lambda}_{h,k}^{-1}}\\
&+\left\|\sum_{i=1}^{k-1}\tilde{\tilde{\sigma}}_{i,h}^{-2}\phi\left(s_h^i,a_h^i\right)\left(\Delta\hat{\tilde{V}}_{k,h+1}(s_{h+1}^i)-\mathbb{P}_h\Delta\hat{\tilde{V}}_{k,h+1}(s_h^i,a_h^i)\right)\right\|_{\tilde{\Lambda}_{h,k}^{-1}}
\end{aligned}$$

where we define $\Delta\hat{\tilde{V}}_{k,h+1} = \hat{\tilde{V}}_{k,h+1} - V_{h+1}^*$. For the first term, we use the result from Zhou and Gu (Lemma G.8) where

$$x_i = \tilde{\tilde{\sigma}}_{i,h}^{-1}\phi(s_h^i,a_h^i)$$

and

$$\eta_i = \mathbb{1}\left\{[\mathbb{V}_hV_{h+1}^*](s_h^i,a_h^i) \le \tilde{\tilde{\sigma}}_{i,h}^2\right\}\left(\tilde{\tilde{\sigma}}_{i,h}^{-1}(V_{h+1}^*(s_{h+1}^i)-[\mathbb{P}_hV_{h+1}^*](s_h^i,a_h^i))\right)$$

Then, we have the following:

$$\begin{aligned}
\|x_i\|_2 = \left\|\tilde{\tilde{\sigma}}_{i,h}^{-1}\phi(s_h^i,a_h^i)\right\|_2 &\le \frac{\|\phi(s_h^i,a_h^i)\|_2}{\sqrt{H}} \le \frac{1}{\sqrt{H}},\\
\mathbb{E}[\eta_i|\mathcal{F}_i] = 0, \quad |\eta_t| &\le \left|\left(\tilde{\tilde{\sigma}}_{i,h}^{-1}(V_{h+1}^*(s_{h+1}^i)-[\mathbb{P}_hV_{h+1}^*](s_h^i,a_h^i))\right)\right| \le \sqrt{H},\\
\mathbb{E}[\eta_i^2|\mathcal{F}_i] = \mathbb{E}\left[\mathbb{1}\left\{[\mathbb{V}_hV_{h+1}^*](s_h^i,a_h^i) \le \tilde{\tilde{\sigma}}_{i,h}^2\right\}\cdot\tilde{\tilde{\sigma}}_{i,h}^{-2}[\mathbb{V}_hV_{h+1}^*](s_h^i,a_h^i)\right] &\le 1,\\
\max_i\left\{|\eta_i|\cdot\min\{1,\|x_i\|_{\tilde{\Lambda}_{i,h}^{-1}}\}\right\} \le 2H\tilde{\tilde{\sigma}}_{i,h}^{-1}\|x_i\|_{\tilde{\Lambda}_{i,h}^{-1}} &\le \sqrt{d}.
\end{aligned}$$

Thus, with probability at least $1-\delta/H$, for all $k\in[K]$, we have

$$\left\|\sum_{i=1}^{k-1}x_i\eta_i\right\|_{\tilde{\Lambda}_{k,h}^{-1}} \le O\left(\sqrt{d\log^2\left(1+\left(\frac{HK^4L^2d}{\delta\tilde{\lambda}_\Lambda}\right)\right)}\right).$$

In addition, on the event $\tilde{\mathcal{E}}_{h+1}$ and $\mathcal{E}$, according to Lemma B.2, we have

$$\tilde{\bar{\sigma}}_{k,h}^2 \geq [\overline{\mathbb{V}}_{k,h}\hat{\tilde{V}}_{k,h+1}](s_h^k, a_h^k) + E_{k,h} + D_{k,h} \geq [\mathbb{V}_h V_{h+1}^*](s_h^k, a_h^k),$$

which further implies that

$$\left\|\sum_{i=1}^{k-1} \tilde{\bar{\sigma}}_{i,h}^{-2}\phi\left(s_h^i, a_h^i\right)\left(V_{h+1}^*(s_{h+1}^i) - \mathbb{P}_h V_{h+1}^*(s_h^i, a_h^i)\right)\right\|_{\tilde{\Lambda}_{h,k}^{-1}} = \left\|\sum_{i=1}^{k-1} x_i\eta_i\right\|_{\tilde{\Lambda}_{k,h}^{-1}} \leq O\left(\sqrt{d\log^2\left(1+\left(\frac{HK^4L^2d}{\delta\tilde{\lambda}_\Lambda}\right)\right)}\right).$$

For the second term, we cannot directly use Lemma G.8 since the stochastic noise

$$\Delta\hat{\tilde{V}}_{k,h+1}(s_{h+1}^i) - [\mathbb{P}_h(\Delta\hat{\tilde{V}}_{k,h+1})](s_h^i, a_h^i)$$

is not $\mathcal{F}_{i+1}$ measurable. Thus, we need to use the ε-net covering argument. For each episode i, the value function $V_{i,h}$ belongs to the optimistic value function class $\mathcal{V}$. If we set $\varepsilon = \sqrt{\tilde{\lambda}_\Lambda}/(4H^2d^2K)$, then according to Lemma F.2, the covering entropy for the function class $\mathcal{V} - V_{h+1}^*$ is upper bounded by

$$\log N_\varepsilon \leq O(d^3H^2\log^2(HK^4L^2d/\tilde{\lambda}_\Lambda)).$$

Then for function $\hat{\tilde{V}}_{k,h}$, there must exist a function $\tilde{V}$ in the ε-net, such that

$$\text{dist}(\Delta\hat{\tilde{V}}_{k,h}, \tilde{V}) \leq \varepsilon.$$

Therefore, the variance of function $\tilde{V}$ is upper bounded by

$$\begin{aligned}
[\mathbb{V}_h\tilde{V}](s_h^i, a_h^i) - [\mathbb{V}_h(\Delta\hat{\tilde{V}}_{k,h+1})](s_h^i, a_h^i) &= [\mathbb{P}_h\tilde{V}^2](s_h^i, a_h^i) - [\mathbb{P}_h(\Delta\hat{\tilde{V}}_{k,h+1})^2](s_h^i, a_h^i) \\
&\quad + ([\mathbb{P}_h(\Delta\hat{\tilde{V}}_{k,h+1})](s_h^i, a_h^i))^2 - (\mathbb{P}_h\tilde{V}(s_h^i, a_h^i))^2 \\
&\leq 2\,\text{dist}(\Delta\hat{\tilde{V}}_{k,h}, \tilde{V}) \cdot \max_{s'}|\Delta\hat{\tilde{V}}_{k,h+1} + \tilde{V}|(s') \\
&\leq 4H \cdot \text{dist}(\Delta\hat{\tilde{V}}_{k,h}, \tilde{V}) \\
&\leq \frac{1}{d^2}
\end{aligned}$$

where the first inequality holds due to the definition of distance between different functions, the third inequality holds since $|\Delta V_{k,h+1}(s') + \tilde{V}(s')| \leq 2H$, and the last inequality holds due to the definition of ε-net. Again, we make use of Lemma G.8 with the following:

$$x_i = \tilde{\bar{\sigma}}_{i,h}^{-1}\phi(s_h^i, a_h^i)$$

and

$$\eta_i = 1\{[\mathbb{V}_h\tilde{V}](s_h^i, a_h^i) \leq \tilde{\bar{\sigma}}_i^2/(d^3H)\} \cdot \tilde{\bar{\sigma}}_i^{-1}(\tilde{V}(s_{h+1}^i) - [\mathbb{P}_h\tilde{V}](s_h^i, a_h^i)).$$

Therefore, for x_t, η_t, we have the following property:

$$\begin{aligned}
\|x_i\|_2 = \|\tilde{\bar{\sigma}}_{i,h}^{-1}\phi(s_h^i, a_h^i)\|_2 \leq \|\phi(s_h^i, a_h^i)\|_2/\sqrt{H} \leq 1/\sqrt{H}, \\
\mathbb{E}[\eta_i \mid \mathcal{F}_i] = 0, \quad |\eta_t| \leq \left|\tilde{\bar{\sigma}}_i^{-1}(V_h^*(s_{h+1}^i) - [\mathbb{P}_h\tilde{V}_{h+1}](s_h^i, a_h^i))\right| \leq \sqrt{H}, \\
\mathbb{E}[\eta_i^2 \mid \mathcal{F}_i] = \mathbb{E}\left[1\{[\mathbb{V}_h\tilde{V}](s_h^i, a_h^i) \leq \tilde{\bar{\sigma}}_i^2/(d^3H)\} \cdot \tilde{\bar{\sigma}}_i^{-2}[\mathbb{V}_h\tilde{V}](s_h^i, a_h^i)\right] \leq \frac{1}{d^3H}, \\
\max_i\left\{|\eta_i| \cdot \min\{1, \|x_i\|_{\Sigma_{i,h}^{-1}}\}\right\} \leq 2H\tilde{\bar{\sigma}}_i^{-1}\|x_i\|_{\tilde{\Lambda}_{i,h}^{-1}} \leq \frac{1}{d^3H}.
\end{aligned}$$

After taking a union bound over the ε-net, with probability at least $1-\delta$, we have

$$\left\|\sum_{i=1}^{k-1} x_i\eta_i\right\|_{\tilde{\Lambda}_{k,h}^{-1}} \leq O\left(\sqrt{d\log^2\left(1+\left(\frac{HK^4L^2d}{\delta\tilde{\lambda}_\Lambda}\right)\right)}\right).$$

In addition, on the event $\tilde{E}_{h+1}$ and E, according to Lemmas B.2 and B.3, we have

$$\tilde{\overline{\sigma}}^2_{i,h} \geq [\overline{\mathbb{V}}_{i,h}\hat{\tilde{V}}_{i,h+1}](s^k_h, a^k_h)+E_{i,h}+D_{i,h}+H \geq D_{i,h}+H \geq d^3H[\mathbb{V}_h(\Delta\hat{\tilde{V}}_{k,h+1})](s^i_h, a^i_h)+H \geq d^3H[\mathbb{V}_h\tilde{V}](s^i_h, a^i_h).$$

Denote $\bar{V} = \Delta\hat{\tilde{V}}_{k,h+1} - \tilde{V}$. Then,

$$\begin{aligned}
&\left\|\sum_{i=1}^{k-1}\tilde{\bar{\sigma}}^{-2}_{i,h}\phi\left(s^i_h, a^i_h\right)\left(\Delta\hat{\tilde{V}}_{k,h+1}(s^i_{h+1}) - \mathbb{P}_h\Delta\hat{\tilde{V}}_{k,h+1}(s^i_h, a^i_h)\right)\right\|_{\tilde{\Lambda}^{-1}_{h,k}} \\
&\leq 2\left\|\sum_{i=1}^{k-1}\tilde{\bar{\sigma}}^{-2}_{i}\phi(s^i_h, a^i_h)(\tilde{V}(s^i_{h+1}) - [\mathbb{P}_h\tilde{V}](s^i_h, a^i_h))\right\|_{\tilde{\Lambda}^{-1}_{k,h}} \\
&\quad + 2\left\|\sum_{i=1}^{k-1}\tilde{\bar{\sigma}}^{-2}_{i}\phi(s^i_h, a^i_h)(\bar{V}(s^i_{h+1}) - [\mathbb{P}_h\bar{V}](s^i_h, a^i_h))\right\|_{\tilde{\Lambda}^{-1}_{k,h}} \\
&\leq 2\left\|\sum_{i=1}^{k-1}x_i\eta_i\right\|_{\tilde{\Lambda}^{-1}_{k,h}} + \frac{8\varepsilon^2k^2}{\lambda} \\
&\leq O\left(\sqrt{d\log^2\left(1+\left(\frac{HK^4L^2d}{\delta\tilde{\lambda}_\Lambda}\right)\right)}\right).
\end{aligned}$$

where the first inequality holds from $\|a+b\|_\Sigma \leq 2\|a\|_\Sigma + 2\|b\|_\Sigma$, the second inequality holds from $|\overline{V}(s')| \leq \varepsilon$, $\|\phi(s,a)\|_2 \leq 1$, and $\tilde{\Lambda}_{k,h} \succeq \tilde{\lambda}_\Lambda$, $\tilde{\Lambda}^{-1}_{k,h} \preceq 1$ and the last inequality holds with $\varepsilon = \sqrt{\tilde{\lambda}_\Lambda}/(4H^2d^2K)$. Combining these results, we get

$$\left|\hat{\tilde{w}}^\top_{k,h'}\phi(s,a) - [\mathbb{P}_h\hat{\tilde{V}}_{k,h'+1}](s,a)\right| \leq \beta\sqrt{\phi(s,a)^\top\tilde{\Lambda}^{-1}_{k,h'}\phi(s,a)}$$

where

$$\beta = O\left(HL\sqrt{d\tilde{\lambda}_\Lambda} + \sqrt{d\log^2\left(1+\left(\frac{HK^4L^2d}{\delta\tilde{\lambda}_\Lambda}\right)\right)}\right)$$

□

C Estimated Variance and Regret Bound Proofs

We simply state some results derived by [He et al. 2023,]. Our results are largely the same except for factors like ι and unlike [He et al. 2023,], we must retain these terms since $\tilde{\lambda}_\Lambda$ has an upper-bound that is induced from noise added to the privatized estimators and is not a regular constant like λ in regular ridge-regression

Lemma C.1 (Lemma 4.4 From [Zhou and Gu 2022,])**.** *For any parameters $\beta' \geq 1$ and $C \geq 1$, the summation of bonuses is upper bounded by*

$$\sum_{k=1}^{K}\min\left(\beta'\sqrt{\phi(s^h_k, a^h_k)^\top\tilde{\Lambda}^{-1}_{k,h}\phi(s^h_k, a^h_k)}, C\right) \leq 4d^4H^6C\iota+10\beta'd^5H^4\iota+2\beta'\sqrt{2d\iota\sum_{k=1}^{K}(\tilde{\sigma}^2_{k,h}+H)},$$

where $\iota = \log\left(1+\frac{K}{d\tilde{\lambda}_\Lambda}\right)$.

Lemma C.2 (Lemma C.1 From [He et al. 2023,])**.** *Define $\mathcal{E}_1$ as the following event*

$$\begin{aligned}
\mathcal{E}_1 = \Big\{\forall h \in [H],\ &\sum_{k=1}^{K}\sum_{h'=h}^{H}\left[\mathbb{P}_h\big(\widehat{V}_{k,h+1} - \widehat{V}^{\hat{\pi}_k}_{k,h+1}\big)\right](s^k_h, a^k_h) \\
&- \sum_{k=1}^{K}\sum_{h'=h}^{H}\big(\widehat{V}_{k,h+1}(s^k_{h+1}) - \widehat{V}^{\hat{\pi}_k}_{k,h+1}(s^k_{h+1})\big) \leq 2\sqrt{2H^3K\log(H/\delta)}\Big\}.
\end{aligned}$$

Then, $\Pr(\mathcal{E}_1) \geq 1-\delta$. *Furthermore, on the events* $\tilde{\mathcal{E}}$, $\mathcal{E}$, *and* $\mathcal{E}_1$, *for all stages* $h \in [H]$, *the regret in the first* K *episodes is upper bounded by:*

$$\sum_{k=1}^{K}\left(\tilde{\hat{V}}_{k,h}(s_h^k) - \tilde{\hat{V}}_{k,h}^{\tilde{\pi}^k}(s_h^k)\right) \leq 16d^4H^8\iota + 40\beta d^7H^5\iota + 8\beta\sqrt{2dH\iota}\sum_{h=1}^{H}\sum_{k=1}^{K}(\tilde{\sigma}_{k,h}^2 + H) + 4\sqrt{H^3K\log(H/\delta)},$$

and for all stages $h \in [H]$, *we further have:*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{P}_h\left(\tilde{\hat{V}}_{k,h}(s_h^k) - \tilde{\hat{V}}_{k,h}^{\tilde{\pi}^k}(s_h^k)\right)(s_h^k, a_h^k) \leq 16d^4H^9\iota + 40\beta d^7H^6\iota + 8H\beta\sqrt{2dH\iota}\sum_{h=1}^{H}\sum_{k=1}^{K}(\tilde{\sigma}_{k,h}^2 + H) + 4\sqrt{H^5K\log(H/\delta)},$$

where $\iota = \log\left(1 + \frac{K}{d\tilde{\lambda}_\Lambda}\right)$.

Lemma C.3. *Lemma C.2 From [He et al. 2023,] Define* $\mathcal{E}_2$ *as the following event*

$$\mathcal{E}_2 = \Big\{\forall h \in [H],\ \sum_{k=1}^{K}\sum_{h'=h}^{H}\big[\mathbb{P}_h\big(\hat{V}_{k,h+1} - \check{V}_{k,h+1}\big)\big](s_h^k, a_h^k) - \sum_{k=1}^{K}\sum_{h'=h}^{H}\big(\hat{V}_{k,h+1}(s_{h+1}^k) - \check{V}_{k,h+1}(s_{h+1}^k)\big) \leq 2\sqrt{2H^3K\log(H/\delta)}\Big\}.$$

Then, $\Pr(\mathcal{E}_2) \geq 1-\delta$. *On the events* $\tilde{\mathcal{E}}$, $\mathcal{E}$, *and* $\mathcal{E}_2$, *the difference between the optimistic value function* $\tilde{\hat{V}}_{k,h}$ *and the pessimistic value function* $\tilde{\check{V}}_{k,h}$ *is upper bounded by:*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{P}_h\left(\tilde{\hat{V}}_{k,h}(s_h^k) - \tilde{\check{V}}_{k,h+1}\right)(s_h^k, a_h^k) \leq 32d^4H^9\iota + 40(\beta+\widehat{\beta})d^7H^6\iota + 8H(\beta+\widehat{\beta})\sqrt{2dH\iota}\sum_{h=1}^{H}\sum_{k=1}^{K}(\tilde{\sigma}_{k,h}^2 + H) + 4\sqrt{H^5K\log(H/\delta)},$$

where $\iota = \log\left(1 + \frac{K}{d\tilde{\lambda}_\Lambda}\right)$.

Lemma C.4 (Lemma C.3 From [He et al. 2023,])**.** *On the event* $\mathcal{E} \cap \tilde{\mathcal{E}} \cap \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$, *the total estimated variance is upper bounded by:*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\tilde{\sigma}_{k,h}^2 \leq O\left(H^2K + H^{4.5}d^3K^{0.5}L^2\log^{1.5}\left(\frac{HK^4L^2d}{\delta\tilde{\lambda}_\Lambda}\right)\right).$$

Lemma C.5. *For any linear MDP* $\mathcal{M}$, *if we set the confidence radii* $\widehat{\beta}$, $\check{\beta}$, $\overline{\beta}$ *as follows:*

$$\widehat{\beta} = \check{\beta} = O\left(HL\sqrt{d\tilde{\lambda}_\Lambda} + \sqrt{d^3H^2\log^2\left(\frac{HK^4L^2d}{\delta\tilde{\lambda}_\Lambda}\right)}\right),$$

$$\bar{\beta} = O\left(H^2L^2\sqrt{d\tilde{\lambda}_\Lambda} + \sqrt{d^3H^4\log^2\left(\frac{HK^4L^2d}{\delta\tilde{\lambda}_\Lambda}\right)}\right).$$

$$\beta = O\left(HL\sqrt{d\tilde{\lambda}_\Lambda} + \sqrt{d\log^2\left(1 + \left(\frac{HK^4L^2d}{\delta\tilde{\lambda}_\Lambda}\right)\right)}\right)$$

then with high probability of at least $1-7\delta$, *the regret of DP-LSVI-UCB*$^{++}$ *is upper bounded as follows:*

$$\textit{Regret}(K) \leq \tilde{O}\left(d\sqrt{H^3K} + \frac{H^{18/4}d^{7/6}K^{1/2}\log(10dKH/\delta)}{\epsilon}\right)$$

In addition, the number of updates for $\tilde{\hat{Q}}_{k,h}$ *and* $\tilde{\check{Q}}_{k,h}$ *is upper bounded by* $O(dH\log(1 + K/d\tilde{\lambda}_\Lambda))$.

Proof of Lemma C.5. On the event $\mathcal{E} \cap \tilde{\mathcal{E}} \cap \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$, the regret is upper bounded by:

$$\begin{aligned}
\text{Regret}(K) &= \sum_{k=1}^{K} \left(V_1^*(s_1^k) - \tilde{\hat{V}}_{k,1}^{\pi_k}(s_1^k) \right) \\
&\leq \sum_{k=1}^{K} \left(\tilde{\hat{V}}_{k,1}(s_1^k) - \tilde{\hat{V}}_{k,1}^{\hat{\pi}^k}(s_1^k) \right) \\
&\leq 16d^4H^8\iota + 40\beta d^7H^5\iota + 8\beta\sqrt{2dH\iota \sum_{h=1}^{H}\sum_{k=1}^{K}(\tilde{\sigma}_{k,h}^2 + H)} + 4\sqrt{H^3K\log(H/\delta)} \\
&\leq \tilde{O}\left(d\sqrt{H^3K} + \frac{H^{18/4}d^{7/6}K^{1/2}\log(10dKH/\delta)}{\epsilon} \right)
\end{aligned}$$

where $\iota = \log(1 + K/(d\tilde{\lambda}_\Lambda))$. The first inequality holds due to optimism (Lemma B.5), the second inequality holds due to Lemma C.3, and the last inequality holds due to the variance bound (Lemma C.4). Since the event $\mathcal{E} \cap \tilde{\mathcal{E}} \cap \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$ holds with probability at least $1 - 7\delta$ holds. In addition, according to Lemma D.2, the number of updates for $\tilde{\hat{Q}}_{k,h}$ and $\tilde{\check{Q}}_{k,h}$ is upper bounded by $O(dH\log(1 + K/\tilde{\lambda}_\Lambda))$. □

D Switching Cost Proof

We first prove a standard determinant upper bound for our privatized Gram matrix $\tilde{\Lambda}$. This will be useful for determining the switching cost

Lemma D.1 (Privatized Determinant Upper Bound (Similar to Lemma C.1 in [Wang et al. 2021,])**.** *Let* $\left\{ \tilde{\Lambda}_{h,k}, (h,k) \in [H] \times [K] \right\}$ *be defined as in Algorithm 1. Then, for all* $h \in [H], k \in [K]$*, we have that* $\det\left(\tilde{\Lambda}_{h,k}\right) \leq (\tilde{\lambda}_\Lambda + (k-1)/d)^d$.

Proof of Lemma D.1. We have that

$$\begin{aligned}
\text{Tr}\left(\tilde{\Lambda}_{h,k}\right) &= \text{Tr}(K_1) + \sum_{i=1}^{k-1} \bar{\sigma}_{i,h}^{-2} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\top \\
&\leq d\tilde{\lambda}_\Lambda + \sum_{i=1}^{k-1} \bar{\sigma}_{i,h}^{-2} \phi(s_h^i, a_h^i)\phi(s_h^i, a_h^i)^\top \\
&\leq d\tilde{\lambda}_\Lambda + \sum_{i=1}^{k-1} ||\phi(s_h^i, a_h^i)||_2 \\
&\leq d\tilde{\lambda}_\Lambda + k - 1
\end{aligned}$$

where the first inequality holds from the fact that for a symmetric matrix A, we have the inequality $\text{Tr}(A) \leq n||A||_2$ and that from the utility analysis (Lemma A.1), $||K_1||_2 \leq \tilde{\lambda}_\Lambda$. The second inequality holds from the fact that $\bar{\sigma}_{i,h}^{-2} \leq 1$, and the last inequality holds from the assumption that $||\phi(s_h^i, a_h^i)||_2 \leq 1$. Now, since we have that $\tilde{\Lambda}_{h,k}$ is positive semi-definite, by the AM-GM inequality, we have

$$\det\left(\tilde{\Lambda}_{h,k}\right) \leq \left(\frac{\text{Tr}\left(\tilde{\Lambda}_{h,k}\right)}{d} \right)^d \leq \left(\tilde{\lambda}_\Lambda + \frac{k-1}{d} \right)^d$$

□

We can finally prove the switching cost of Algorithm 1.

Lemma D.2. *Conditioned on the event that $||K_1||_2 \leq \tilde{\lambda}_\Lambda$ for all $h, k \in [H] \times [K]$, DP-LSVI-UCB^{++} (Algorithm 1) has a global switching cost of atmost $O\left(dH \log(1 + K/d\tilde{\lambda}_\Lambda)\right)$.*

Proof. We denote $k_0 = 0$ and suppose that $\{k_1, \ldots, k_m\}$ be the episodes where our algorithm updates the value function. Then, according to the determinant-based criterion (Line 10), for each episode k_i, there exists an $h \in [H]$ such that $\det(\tilde{\Lambda}_{k,h}) \geq 2\det(\tilde{\Lambda}_{k_{i-1},h})$. Then, due to the utility analysis (Lemma A.1), for $h' \neq h$, we have $\tilde{\Lambda}_{k_i,h'} \succeq \tilde{\Lambda}_{k_{i-1},h'}$. Thus,

$$\prod_{h=1}^{H} \det(\tilde{\Lambda}_{k,h}) \geq 2 \prod_{h=1}^{H} \det(\tilde{\Lambda}_{k_{i-1},h})$$

Applying this across all episodes, we get

$$\prod_{h=1}^{H} \det(\tilde{\Lambda}_{k,h}) \geq 2^m \prod_{h=1}^{H} \det(\tilde{\Lambda}_{k_0,h}) = 2^m \prod_{h=1}^{H} \det(2\tilde{\lambda}_\Lambda I) = 2^m \tilde{\lambda}_\Lambda^{dH}$$

Furthermore, from Lemma D.1

$$\prod_{h=1}^{H} \det(\tilde{\Lambda}_{k,h}) \leq \left(\tilde{\lambda}_\Lambda + \frac{K}{d}\right)^{dH}$$

Combining these two inequalities, we conclude with

$$m \leq O\left(dH \log\left(1 + K/d\tilde{\lambda}_\Lambda\right)\right)$$

□

E Weight Norm Proofs

We prove upper bounds on the optimistic, pessimistic, and squared weight vectors. These will be used in uniform covering arguments which are used in our regret analysis.

Lemma E.1. *For all stages $h \in [H]$ and all episodes $n \in \mathbb{N}$, the norm of the weight vector $\tilde{\widehat{w}}_{k,h}$ can be upper bounded as*

$$\left\|\tilde{\widehat{w}}_{k,h}\right\|_2 \leq HKL\sqrt{\frac{2d}{\tilde{\lambda}_\Lambda}}$$

Proof of Lemma E.1. First, recall that by definition we have

$$\tilde{\widehat{w}}_{k,h} = \tilde{\Lambda}_{k,h}^{-1}\left[\sum_{i=1}^{k-1} \tilde{\widehat{\sigma}}_{i,h}^{-2} \phi\left(s_h^i, a_h^i\right) \tilde{\widehat{V}}_{k,h+1}(s_{h+1}^i) + \phi_1\right]$$

Then, we have

$$\begin{aligned}
\left\|\tilde{\widehat{w}}\right\|_2^2 &= \left\|\tilde{\Lambda}_{k,h}^{-1}\left[\sum_{i=1}^{k-1} \tilde{\widehat{\sigma}}_{i,h}^{-2} \phi\left(s_h^i, a_h^i\right) \tilde{\widehat{V}}_{k,h+1}(s_{h+1}^i) + \phi_1\right]\right\|_2^2 \\
&\leq k \sum_{i=1}^{k-1} \left\|\tilde{\Lambda}_{k,h}^{-1} \tilde{\widehat{\sigma}}_{i,h}^{-2} \phi\left(s_h^i, a_h^i\right) \tilde{\widehat{V}}_{k,h+1}(s_{h+1}^i) + \tilde{\Lambda}_{k,h}^{-1}\phi_1\right\|_2^2 \\
&\leq k \sum_{i=1}^{k-1} \left\|\tilde{\Lambda}_{k,h}^{-1} \tilde{\widehat{\sigma}}_{i,h}^{-2} \phi\left(s_h^i, a_h^i\right) \tilde{\widehat{V}}_{k,h+1}(s_{h+1}^i)\right\|_2^2 + k \sum_{i=1}^{k-1} \left\|\tilde{\Lambda}_{k,h}^{-1}\phi_1\right\|_2^2 \\
&\leq \frac{kH^2}{\tilde{\lambda}_\Lambda} \sum_{i=1}^{k-1} \tilde{\widehat{\sigma}}_{i,h}^{-2} \phi\left(s_h^i, a_h^i\right)^\top \tilde{\Lambda}_{k,h}^{-1} \phi\left(s_h^i, a_h^i\right) + \frac{k^2L^2}{\tilde{\lambda}_\Lambda^2} \\
&\leq \frac{kH^2}{\tilde{\lambda}_\Lambda} \text{trace}\left(\tilde{\Lambda}_{k,h}^{-1} \sum_{i=1}^{k-1} \tilde{\widehat{\sigma}}_{i,h}^{-2} \phi\left(s_h^i, a_h^i\right)^\top \phi\left(s_h^i, a_h^i\right)\right) + \frac{k^2L^2}{\tilde{\lambda}_\Lambda^2}
\end{aligned}$$

where the first inequality holds from Cauchy-Schwartz, the second inequality holds from triangle inequality, and the third inequality holds from the fact that $\tilde{\hat{V}}_{k,h+1} \leq H$, $\left\|\tilde{\Lambda}_{k,h}^{-1}\right\|_2 \leq \frac{1}{\tilde{\lambda}_\Lambda}$, and $\|\phi_1\|_2 \leq L$ from the utility analysis (Lemma A.1). Now, we assume the eigen-decomposition of matrix $\sum_{i=1}^{k-1} \tilde{\bar{\sigma}}_{i,h}^{-2} \phi\left(s_h^i, a_h^i\right)^\top \phi\left(s_h^i, a_h^i\right)$ is $Q^\top \Sigma Q$. Then, we have

$$\begin{aligned}
\operatorname{trace}\left(\tilde{\Lambda}_{k,h}^{-1} \sum_{i=1}^{k-1} \tilde{\bar{\sigma}}_{i,h}^{-2} \phi\left(s_h^i, a_h^i\right)^\top \phi\left(s_h^i, a_h^i\right)\right) &= \operatorname{trace}\left(\left(Q^\top \Sigma Q + 2\tilde{\lambda}_\Lambda I_d\right)^{-1} Q^\top \Sigma Q\right) \\
&= \operatorname{trace}\left(\left(\Sigma + 2\tilde{\lambda}_\Lambda I_d\right)^{-1} \Sigma\right) \\
&= \sum_{i=1}^{d} \frac{\sigma_i}{\sigma_i + 2\tilde{\lambda}_\Lambda I_d} \\
&\leq d
\end{aligned}$$

Thus, putting these together, we get

$$\left\|\tilde{\hat{w}}_{k,h}\right\|_2^2 \leq \frac{kH^2 d}{\tilde{\lambda}_\Lambda} + \frac{k^2 L^2}{\tilde{\lambda}_\Lambda^2} \leq \frac{2k^2 H^2 L^2 d}{\tilde{\lambda}_\Lambda}$$

□

The same analysis holds for the pessimistic weight vector $\tilde{\check{w}}$

Lemma E.2. *For all stages $h \in [H]$ and all episodes $n \in \mathbb{N}$, the norm of the weight vector $\tilde{\check{w}}_{k,h}$ can be upper bounded as*

$$\left\|\tilde{\check{w}}_{k,h}\right\|_2 \leq HKL\sqrt{\frac{2d}{\tilde{\lambda}_\Lambda}}$$

Proof of Lemma E.2. The proof is exactly the same as Lemma E.1 except we use the pessimistic value function class $\check{\mathcal{V}}_h$. □

Likewise, using similar analysis as above, we can also bound the weight vector $\tilde{\overline{w}}_{k,h}$.

Lemma E.3. *For all stages $h \in [H]$ and all episodes $n \in \mathbb{N}$, the norm of the weight vector $\tilde{\overline{w}}_{k,h}$ can be upper bounded as*

$$\left\|\tilde{\overline{w}}_{k,h}\right\|_2 \leq H^2 KL\sqrt{\frac{2d}{\tilde{\lambda}_\Lambda}}$$

Proof of Lemma E.3. The proof is exactly the same as Lemma E.1 except we use the pessimistic value function class $\widehat{\mathcal{V}}_h^2$. □

F Covering Argument Results

Lemma F.1 (Lemma D.5 from [Jin et al. 2020,])**.** *For a Euclidean ball with radius R in $\mathbb{R}^d$, the ε-covering number of this ball is upper bounded by*

$$(1 + 2R/\varepsilon)^d.$$

With the help of Lemma F.5, the covering number $\mathcal{N}_\varepsilon$ of optimistic function class $\widehat{\mathcal{V}}_h$ can be upper bounded by the following lemma:

Lemma F.2 (Lemma F.6 from [He et al. 2023,])**.** *For optimistic function class $\widehat{\mathcal{V}}_h$,*

$$\widehat{\mathcal{V}}_h = \left\{ V \;\middle|\; V(\cdot) = \max_a \min_{1 \leq i \leq l} \min\left(H, r_h(\cdot, a) + w_i^\top \phi(\cdot, a) + \beta\sqrt{\phi(\cdot, a)^\top \tilde{\Lambda}_i^{-1} \phi(\cdot, a)}\right), \|w_i\| \leq L_1, \tilde{\Lambda}_i \succeq \tilde{\lambda}_\Lambda I \right\}$$

where $l = dH \log\left(1 + K/d\tilde{\lambda}_\Lambda\right)$ *and* $L_1 = HKL\sqrt{\frac{2d}{\tilde{\lambda}_\Lambda}}$. *Define the distance between two functions* V_1 *and* V_2 *as* $V_1, V_2 \in \widehat{\mathcal{V}}_h$ *as dist*$(V_1, V_2) = \max_s |V_1(s) - V_2(s)|$. *With respect to this distance function, the* ε*-covering number* $\mathcal{N}_\varepsilon$ *of the function class* $\mathcal{V}_h$ *can be upper bounded by*

$$\log \mathcal{N}_\varepsilon \leq dl \log(1 + 4L_1/\varepsilon) + d^2 l \log\left(1 + 8\sqrt{d}\beta^2 / \left(\tilde{\lambda}_\Lambda \varepsilon^2\right)\right).$$

Proof of Lemma F.2. By letting $\Sigma = \beta^2 \left(\tilde{\Lambda}\right)^{-1}$, we can reparametrize the function class $\widehat{\mathcal{V}}_h$ as

$$\widehat{\mathcal{V}}_h = \left\{ V \;\middle|\; V(\cdot) = \max_a \min_{1 \leq i \leq l} \min\left(H, r_h(\cdot, a) + w_i^\top \phi(\cdot, a) + \sqrt{\phi(\cdot, a)^\top \Sigma \phi(\cdot, a)}\right), \|w_i\| \leq L_1, \Sigma \succeq \beta^2 \tilde{\lambda}_\Lambda I \right\}$$

For any two functions $V_1, V_2 \in \widehat{\mathcal{V}}_h$, let them take the form as seen above. Then, since $\min\{H, \cdot\}$, $\min_{1 \leq i \leq l}$, and $\max_a$ are contraction maps, we have

$$\begin{aligned}
\operatorname{dist}(V_1, V_2) &= \max_{s \in \mathcal{S}} |V_1(s) - V_2(s)| \\
&\leq \max_{1 \leq i \leq l, s \in \mathcal{S}, a \in \mathcal{A}} \left| w_{1,i}^\top \phi(\cdot, a) + \sqrt{\phi(\cdot, a)^\top \Sigma_{1,i} \phi(\cdot, a)}\right) - w_{2,i}^\top \phi(\cdot, a) + \sqrt{\phi(\cdot, a)^\top \Sigma_{2,i} \phi(\cdot, a)}\right) \Bigg| \\
&\leq \max_{1 \leq i \leq l, s \in \mathcal{S}, a \in \mathcal{A}} \left| \left(w_{1,i} - w_{2,i}\right) \phi(s, a) \right| + \max_{1 \leq i \leq l, s \in \mathcal{S}, a \in \mathcal{A}} \left| \sqrt{\phi(s, a)^\top \left(\Sigma_{1,i} - \Sigma_{2,i}\right) \phi(s, a)} \right| \\
&\leq \max_{1 \leq i \leq l} \left\| w_{1,i} - w_{2,i} \right\|_2 + \max_{1 \leq i \leq l} \sqrt{\left\| \Sigma_{1,i} - \Sigma_{2,i} \right\|_F}
\end{aligned}$$

where the first inequality holds due to the contraction property, the second inequality holds due to the fact that $\max_x |f(x) + g(x)| \leq \max_x |f(x)| + \max_x |g(x)|$ and $\left|\sqrt{x} - \sqrt{y}\right| \leq \sqrt{|x - y|}$, and the last inequality holds from $\|\phi(s, a\|_2 \leq 1$. Now, let $\mathcal{C}_w$ be a $\varepsilon/2$ covering net of $\left\{w \in \mathbb{R}^d \mid \|w\|_2 \leq L_1\right\}$ and let $\mathcal{C}_\Sigma$ be a $\varepsilon^2/4$ covering net of $\left\{\Sigma \in \mathbb{R}^{d \times d} \mid \|\Sigma\|_F \leq d^{1/2} \beta^2 \tilde{\lambda}_\Lambda^{-1}\right\}$. By Lemma F.1, we know

$$|\mathcal{C}_w| \leq (1 + 4L/\varepsilon)^d, \; |\mathcal{C}_\Sigma| \leq \left(1 + 8d^{1/2}\beta^2 / \left(\tilde{\lambda}_\Lambda \varepsilon^2\right)\right)^{d^2}$$

We know that for any $V_1, V_2 \in \widehat{\mathcal{V}}_h$, there exists $w_1, w_2 \in \mathcal{C}_w$ and $\Sigma_1, \Sigma_2 \in \mathcal{C}_\Sigma$ such that $\operatorname{dist}(V_1, V_2) \leq \varepsilon$. Thus, this means that the covering number $|N_\varepsilon| \leq |\mathcal{C}_w|^l |\mathcal{C}_\Sigma|^l$. Thus, taking logs, we get

$$\log \mathcal{N}_\varepsilon \leq dl \log(1 + 4L_1/\varepsilon) + d^2 l \log\left(1 + 8\sqrt{d}\beta^2 / \left(\tilde{\lambda}_\Lambda \varepsilon^2\right)\right).$$

□

Likewise, we can also upper bound the covering number of the pessimistic function class $\check{\mathcal{V}}_h$

Lemma F.3 (Lemma F.7 from [He et al. 2023,])**.** *For pessimistic function class* $\check{\mathcal{V}}_h$*,*

$$\check{\mathcal{V}}_h = \left\{ V \;\middle|\; V(\cdot) = \max_a \max_{1 \leq i \leq l} \max\left(H, r_h(\cdot, a) + w_i^\top \phi(\cdot, a) - \beta\sqrt{\phi(\cdot, a)^\top \tilde{\Lambda}_i^{-1} \phi(\cdot, a)}\right), \|w_i\| \leq L_1, \tilde{\Lambda}_i \succeq \tilde{\lambda}_\lambda I \right\}$$

where $l = dH \log\left(1 + K/d\tilde{\lambda}_\Lambda\right)$ *and* $L_1 = HKL\sqrt{\frac{2d}{\tilde{\lambda}_\Lambda}}$. *Define the distance between two functions* V_1 *and* V_2 *as* $V_1, V_2 \in \check{\mathcal{V}}_h$ *as dist*$(V_1, V_2) = \max_s |V_1(s) - V_2(s)|$. *With respect to this distance function, the* ε*-covering number* $\mathcal{N}_\varepsilon$ *of the function class* $\check{\mathcal{V}}_h$ *can be upper bounded by*

$$\log \mathcal{N}_\varepsilon \leq dl \log(1 + 4L_1/\varepsilon) + d^2 l \log\left(1 + 8\sqrt{d}\beta^2 / \left(\tilde{\lambda}_\Lambda \varepsilon^2\right)\right).$$

Now that we have these results, the only result we require is an upper bound on the covering number of the optimistic value function class squared. This result is provided below

Lemma F.4 (Lemma F.7 from [He et al. 2023,]). *For the squared function class $\widehat{\mathcal{V}}_h^2$, we define the distance between two functions V_1^2 and V_2^2 in $\widehat{\mathcal{V}}_h^2$ as:*

$$dist(V_1^2, V_2^2) = \max_s \left|V_1^2(s) - V_2^2(s)\right|.$$

With respect to this distance function, the ε-covering number N_ε of the function class $\widehat{\mathcal{V}}_h^2$ can be upper bounded by:

$$\log N_\varepsilon \leq dl \log\left(1 + 8HL_1/\varepsilon\right) + d^2 l \log\left(1 + 32\sqrt{d}H^2\beta^2 / \left(\tilde{\lambda}_\Lambda \varepsilon^2\right)\right).$$

where $l = dH \log\left(1 + K/d\tilde{\lambda}_\Lambda\right)$ and $L_1 = HKL\sqrt{\frac{2d}{\tilde{\lambda}_\Lambda}}$.

G Auxiliary Results

Lemma G.1 (Lemma G.1 From [He et al. 2023,]). *For any stage $h \in [H]$ in a linear MDP and any bounded function $V : S \to [0, B]$, there always exists a vector $w \in \mathbb{R}^d$ such that for all state-action pairs $(s, a) \in S \times A$, we have*

$$[\mathbb{P}_h V](s, a) = w^\top \phi(s, a),$$

where $\|w\|_2 \leq B\sqrt{d}$.

Proof of Lemma G.1. By assumption of the linear MDP setting, we have

$$\begin{aligned}
[\mathbb{P}_h V](s, a) = \int \mathbb{P}_h(s'|s, a)V(s')\, ds' &= \int \phi(s, a)^\top V(s')\, d\theta_h(s') \\
&= \phi(s, a)^\top \int V(s')\, d\theta_h(s') \\
&= \phi(s, a)^\top w,
\end{aligned}$$

where we set $w = \int V(s')\, d\theta_h(s')$. Additionally, the norm of w is upper bounded by $\left\|\int V(s')\, d\theta_h(s')\right\| \leq \max_{s'} V(s') \cdot \sqrt{d} = B\sqrt{d}$. □

Lemma G.2 (Azuma-Hoeffding Inequality, [Cesa-Bianchi and Lugosi 2006,]). *Let $\{x_i\}_{i=1}^n$ be a martingale difference sequence with respect to a filtration $\{\mathcal{G}_i\}$ satisfying $|x_i| \leq M$ for some constant M, x_i is $\mathcal{G}_{i+1}$-measurable, and $\mathbb{E}[x_i|\mathcal{G}_i] = 0$. Then, for any $0 < \delta < 1$, with probability at least $1 - \delta$, we have:*

$$\sum_{i=1}^n x_i \leq M\sqrt{2n \log(1/\delta)}.$$

Lemma G.3 (Lemma 11 in [Abbasi-Yadkori et al. 2011,]). *Let $\{x_k\}_{k=1}^K$ be a sequence of vectors in $\mathbb{R}^d$, and let Σ_0 be a $d \times d$ positive definite matrix. Define $\Sigma_k = \Sigma_0 + \sum_{i=1}^k x_i x_i^\top$. Then, we have:*

$$\sum_{i=1}^k \min\{1, x_i^\top \Sigma_{i-1}^{-1} x_i\} \leq 2 \log\left(\frac{\det \Sigma_k}{\det \Sigma_0}\right).$$

In addition, if $\|x_i\|_2 \leq L$ for all $i \in [K]$, then:

$$\sum_{i=1}^k \min\{1, x_i^\top \Sigma_{i-1}^{-1} x_i\} \leq 2 \log\left(\frac{\det \Sigma_k}{\det \Sigma_0}\right) \leq 2\left(d \log\left(\frac{\mathrm{trace}(\Sigma_0) + kL^2}{d}\right) - \log\det \Sigma_0\right).$$

Lemma G.4 (Lemma 12 in [Abbasi-Yadkori et al. 2011,]). *Suppose $A, B \in \mathbb{R}^{d\times d}$ are two positive definite matrices satisfying $A \preceq B$. Then, for any $x \in \mathbb{R}^d$:*

$$\|x\|_A \leq \|x\|_B \cdot \sqrt{\frac{\det(A)}{\det(B)}}.$$

Lemma G.5 (Theorem 1 in [Abbasi-Yadkori et al. 2011,]). *Let $\{\mathcal{F}_t\}_{t=0}^{\infty}$ be a filtration. Let $\{\eta_t\}_{t=1}^{\infty}$ be a real-valued stochastic process such that η_t is $\mathcal{F}_t$-measurable and η_t is conditionally R-sub-Gaussian for some $R \geq 0$, i.e.,*

$$\forall \lambda \in \mathbb{R}, \quad \mathbb{E}\left[e^{\lambda \eta_t} \mid \mathcal{F}_{t-1}\right] \leq \exp\left(\frac{\lambda^2 R^2}{2}\right).$$

Let $\{x_t\}_{t=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process such that x_t is $\mathcal{F}_{t-1}$-measurable. Assume that Z is a $d \times d$ positive definite matrix. For any $k \geq 0$, define

$$Z_k = Z + \sum_{s=1}^{t} X_s X_s^\top$$

Then, for any $\delta > 0$, with probability at least $1 - \delta$, for all $t \geq 0$,

$$\|\sum_{i=1}^{k} x_i \eta_i\|_{Z_k^{-1}}^2 \leq 2R^2 \log\left(\frac{\det(Z_k)^{1/2} \det(Z)^{-1/2}}{\delta}\right).$$

Lemma G.6 (Confidence Ellipsoid, Theorem 2 in [Abbasi-Yadkori et al. 2011,]). *Let $\{\mathcal{G}_k\}_{k\geq 1}$ be a filtration, and $\{x_k, \eta_k\}_{k\geq 1}$ be a stochastic process such that $x_k \in \mathbb{R}^d$ is $\mathcal{G}_k$-measurable and $\eta_k \in \mathbb{R}$ is $\mathcal{G}_{k+1}$-measurable. Let $L, \sigma, \Sigma, \varepsilon > 0$, and $\mu^* \in \mathbb{R}^d$. For $k \geq 1$, let $y_k = \langle \mu^*, x_k \rangle + \eta_k$, and suppose that η_k, x_k satisfy:*

$$\mathbb{E}[\eta_k | \mathcal{G}_k] = 0, \quad |\eta_k| \leq R, \quad \|x_k\|_2 \leq L.$$

Define $Z_k = 2\tilde{\lambda}_\Lambda I + \sum_{i=1}^{k} x_i x_i^\top + K_1$, $b_k = \sum_{i=1}^{k} y_i x_i$, $\mu_k = Z_k^{-1} b_k$, and:

$$\beta_k = R\sqrt{d \log\left(\frac{1 + kL^2/\tilde{\lambda}_\Lambda}{\delta}\right)}.$$

Then, for any $0 < \delta < 1$, with probability at least $1 - \delta$, we have:

$$\forall k \geq 1, \quad \left\|\sum_{i=1}^{k} x_i \eta_i\right\|_{Z_k^{-1}} \leq \beta_k, \quad \|\mu_k - \mu^*\|_{Z_k} \leq \beta_k + \sqrt{\lambda}\|\mu^*\|_2.$$

Proof of Lemma G.6. We will prove the following determinant-trace inequality. The result will then hold by applying Lemma G.5

Lemma G.7 (Determinant-Trace Inequality). *Suppose $x_1, x_2, \ldots, x_K \in \mathbb{R}^d$ and for any $1 \leq k \leq K$, $\|x_k\|_2 \leq L$. Let $Z_k = 2\tilde{\lambda}_\Lambda I + \sum_{k=1}^{K} x_k x_k^\top + K_1$ where $\|K_1\|_2 \leq \tilde{\lambda}_\Lambda$. Then,*

$$\det(Z_k) \leq (3\tilde{\lambda}_\Lambda + kL^2/d)^d.$$

Proof of Lemma G.7. Let $\alpha_1, \alpha_2, \ldots, \alpha_d$ denote the eigenvalues of Z_k. Recall that from the utility analysis (Lemma A.1), by construction, Z_k must be positive-definite. Then, notice that $\det(Z_k) = \prod_{i=1}^{d} \alpha_k$ and $\operatorname{trace}(Z_k) = \sum_{i=1}^{d} \alpha_k$. By the AM-GM inequality

$$\sqrt[d]{\alpha_1 \ldots \alpha_d} \leq \frac{1}{d} \sum_{i=1}^{d} \alpha_i$$

Thus, we have that $\det(Z_k) \leq (\operatorname{trace}(Z_k)/d)^d$. Furthermore, notice that

$$\begin{aligned}
\operatorname{trace}(Z_k) &= \operatorname{trace}\left(2\tilde{\lambda}_\Lambda I\right) + \operatorname{trace}\left(\sum_{k=1}^{K} x_s x_s^\top\right) + \operatorname{trace}(K_1) \\
&\leq 3d\tilde{\lambda}_\Lambda + KL^2
\end{aligned}$$

where the inequality holds from the assumption that $\|x_k\|_2 \leq L$ and $\|K_1\|_2 \leq \tilde{\lambda}_\Lambda$ from the utility analysis. Thus, putting these together, we get the claim □

We now use the above result. From Lemma G.5, we have that

$$\|\sum_{i=1}^{k} x_i\eta_i\|^2_{Z_k^{-1}} \leq 2R^2 \log\left(\frac{\det(Z_k)^{1/2}\det(Z)^{-1/2}}{\delta}\right).$$

In our case, we have $Z = 2\tilde{\lambda}_\Lambda I$. Utilizing our determinant upper bound and the fact that $\det(Z) = \left(2\tilde{\lambda}_\Lambda\right)^d$, we have

$$\begin{aligned}\log\left(\frac{\det(Z_k)^{1/2}}{\det(Z)^{1/2}}\right) &\leq \log\left(\frac{\left(3\tilde{\lambda}_\Lambda + kL^2/d\right)^{d/2}}{\left(2\tilde{\lambda}_\Lambda\right)^{d/2}}\right) \\ &\leq \frac{d}{2}\log\left(1 + kL^2/\tilde{\lambda}_\Lambda\right)\end{aligned}$$

where the first inequality comes from Lemma G.7 and the last inequality holds just by upper bounding the first constant term in the logarithm. Thus, we get the claim simply by taking square roots. □

Lemma G.8 (Lemma 4.4 in [Zhou and Gu 2022,]). *Let $\{\sigma_k, \hat{\beta}_k\}_{k\geq 1}$ be a sequence of non-negative numbers, $\alpha, \gamma > 0$, $\{a_k\}_{k\geq 1} \subset \mathbb{R}^d$, and $\|a_k\|_2 \leq A$. Let $\{\bar{\sigma}_k\}_{k\geq 1}$ and $\{\hat{\Sigma}_k\}_{k\geq 1}$ be recursively defined as follows:*

$$\hat{\Sigma}_1 = 2\tilde{\lambda}_\Lambda I, \quad \forall k \geq 1,\ \bar{\sigma}_k = \max\{\sigma_k, \alpha, \gamma\|a_k\|^{1/2}_{\hat{\Sigma}_k^{-1}}\}, \quad \hat{\Sigma}_{k+1} = \hat{\Sigma}_k + a_k a_k^\top/\bar{\sigma}_k^2$$

Let $\iota = \log\left(1 + \frac{KA^2}{d\tilde{\lambda}_\Lambda \alpha^2}\right)$. Then, we have:

$$\sum_{k=1}^{K} \min\{1, \|a_k\|_{\hat{\Sigma}_k^{-1}}\} \leq 2d\iota + 2\gamma^2 d\iota + 2\sqrt{d\iota}\sqrt{\sum_{k=1}^{K}(\sigma_k^2 + \alpha^2)}.$$

Proof of Lemma G.8. We refer readers to Lemma 4.4 in [Zhou and Gu 2022,] for further details. Our proofs are identical except for our usage of Lemma G.6 which is why our ι term is different. □

Lemma G.9 (Theorem 4.3 in [Zhou and Gu 2022,]). *Let $\{\mathcal{G}_k\}_{k=1}^\infty$ be a filtration, and let $\{(x_k, \eta_k)\}_{k\geq 1}$ be a stochastic process such that $x_k \in \mathbb{R}^d$ is $\mathcal{G}_k$-measurable and $\eta_k \in \mathbb{R}$ is $\mathcal{G}_{k+1}$-measurable. Let $L, \sigma > 0$, and $\mu^* \in \mathbb{R}^d$. For $k \geq 1$, define $y_k = \langle \mu^*, x_k\rangle + \eta_k$. Suppose that η_k, x_k also satisfy*

$$\mathbb{E}[\eta_k \mid \mathcal{G}_k] = 0, \quad \mathbb{E}[\eta_k^2 \mid \mathcal{G}_k] \leq \sigma^2, \quad |\eta_k| \leq R, \quad \|x_k\|_2 \leq L.$$

For $k \geq 1$, let

$$Z_k = \lambda I + \sum_{i=1}^{k} x_i x_i^\top, \quad b_k = \sum_{i=1}^{k} y_i x_i, \quad \mu_k = Z_k^{-1} b_k,$$

and

$$\beta_k = \widetilde{O}\big(\sigma\sqrt{d} + \max_{1\leq i\leq k} |\eta_i| \min\{1, \|x_i\|_{Z_{i-1}^{-1}}\}\big).$$

Then, for any $0 < \delta < 1$, with probability at least $1 - \delta$, for all $k \in [K]$, we have

$$\left\|\sum_{i=1}^{k} x_i\eta_i\right\|_{Z_k^{-1}} \leq \beta_k, \quad \textit{and} \quad \|\mu_k - \mu^*\|_{Z_k} \leq \beta_k + \sqrt{\lambda}\|\mu^*\|_2.$$