# Improved Regret for Differentially Private Exploration in Linear MDP

**Dung Daniel Ngo** [*1] **Giuseppe Vietri** [*1] **Zhiwei Steven Wu** [2]

## Abstract

We study privacy-preserving exploration in sequential decision-making for environments that rely on sensitive data such as medical records. In particular, we focus on solving the problem of reinforcement learning (RL) subject to the constraint of (joint) differential privacy in the linear MDP setting, where both dynamics and rewards are given by linear functions. Prior work on this problem due to (Luyo et al., 2021) achieves a regret rate that has a dependence of $O(K^{3/5})$ on the number of episodes $K$. We provide a private algorithm with an improved regret rate with an optimal dependence of $O(\sqrt{K})$ on the number of episodes. The key recipe for our stronger regret guarantee is the *adaptivity* in the policy update schedule, in which an update only occurs when sufficient changes in the data are detected. As a result, our algorithm benefits from *low switching cost* and only performs $O(\log(K))$ updates, which greatly reduces the amount of privacy noise. Finally, in the most prevalent privacy regimes where the privacy parameter $\epsilon$ is a constant, our algorithm incurs negligible privacy cost—in comparison with the existing non-private regret bounds, the additional regret due to privacy appears in lower-order terms.

## 1. Introduction

Many real world machine learning (ML) systems operate under the setting of *interactive learning*, where learning algorithms interact with users and collect feedback from them over time. In domains such as personalized medicine where ML algorithms rely on sensitive data, the ability to protect users' data privacy has become increasingly critical. In order to provide rigorous and formal privacy guarantees, there has been a growing and longstanding literature on designing ML algorithms subject to the constraint of *differential privacy* (DP) (Dwork et al., 2006). While the vast majority of the DP ML literature has focused on the setting of supervised learning, there has been significantly less development on interactive learning, especially reinforcement learning (RL).

In this paper, we advance a recent line of work starting from Vietri et al. (2020) that provides RL algorithms with provable privacy guarantees and performance bounds. Concretely, we consider a setting of episodic RL where an agent interacts with $K$ users that arrive sequentially over $K$ episodes. In each episode $k$, the agent interacts with user $k$ over a fixed horizon of $H$ time steps. At each time step in the episode, the current user reveals their state to the agent, then the agent provides the user with a recommended action to take, which subsequently generates a reward received by the user. The goal of the agent is to maximize the cumulative reward over all users, or equivalently, minimize the regret with respect to the optimal policy. In this model, the sequence of states and rewards of each user is considered sensitive information. While each user may be willing to share such information to the agent in exchange for services or recommendations, there is still a risk that the agent inadvertently leaks the user's private information through the interactions with other users.

In order to prevent such privacy risks, existing work on private contextual bandits (Shariff & Sheffet, 2018) and private RL (Vietri et al., 2020) have adopted the notion of *joint differential privacy* (JDP) (Kearns et al., 2014a), a variant of DP that is suitable for sequential learning settings. Informally, JDP requires that for any user $k$, the output information to all other users except $k$ cannot reveal much about $k$'s private data. As a consequence, even if all other users collude (e.g., collectively probing the agent's policies) against $k$, the private information of $k$ is still protected.

Under the constraint of JDP, earlier work by Vietri et al. (2020); Garcelon et al. (2020) focuses on the tabular MDP setting, where the states and actions are discrete and the value functions can be stored in a table. More recent work has considered function approximation for DP RL. In par-

[*]Equal contribution [1] Department of Computer Science & Engineering, University of Minnesota, Minneapolis, MN 55455, USA [2]School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA. Correspondence to: Dung Daniel Ngo <ngo00054@umn.edu>, Giuseppe Vietri <vietr002@umn.edu>, Zhiwei Steven Wu <zstevenwu@cmu.edu>.

*Proceedings of the $39^{th}$ International Conference on Machine Learning*, Baltimore, Maryland, USA, PMLR 162, 2022. 

ticular, Luyo et al. (2021) considers the *linear MDP setting*, in which the transition dynamics and reward function are assumed to be linear. All of these algorithms obtain the JDP guarantee under the so-called *billboard model* (Hsu et al., 2016), in which the RL agent continuously and differentially privately releases a collection of statistics that are sufficient for computing each user's recommended actions when given their own private data. The key step to obtain low regret is to maintain these sufficient statistics at a low privacy costs. To that end, these algorithms in prior work leverage the *binary mechanism* (and its variations) (Dwork et al., 2010; Chan et al., 2011; Shariff & Sheffet, 2018), which can continually release any form of summation or count statistics subject to DP with error only scaling logarithmically in the total number of episodes.

While bringing function approximation to private RL greatly expands its scope and practicality, it also introduces new challenges that the existing privacy techniques based on the binary mechanism cannot handle. In particular, algorithms in the linear MDP setting typically maintain and update the value function $V$, which cannot be re-written as a form of summation statistics on the private data. This barrier has led to a sub-optimal regret in the prior work of (Luyo et al., 2021). Since their algorithm cannot leverage the binary mechanism for tracking the value functions privately, they resorts to applying the Gaussian mechanism repeatedly over an non-adaptive schedule, which incurs a privacy cost that scales with $K^{3/5}$ in their regret. Our work provides a private RL algorithm that circumvents this barrier and achieves a significantly lower privacy cost—in comparison to the non-private bounds, the privacy cost is in lower order.

### 1.1. Our contributions.

We focus on the setting of linear MDP, where there exists a feature map $\phi$ that maps each state-action pair $(x, a)$ to a $d$-dimensional vector $\phi(x, a)$. For each $(x, a)$, we assume that both the reward and transition function are linear in $\phi(x, a)$. Our contributions include the following:

**Our results.** We provide an $(\varepsilon, \delta)$-JDP RL algorithm that achieves the state-of-the-art regret bound of

$$\widetilde{O}\left(\sqrt{d^3H^4K} + \frac{H^3d^{5/4}K^{1/2}}{\epsilon^{1/2}}\right),$$

The best known bound in prior work Luyo et al. (2021) is $\widetilde{O}\left(\sqrt{d^3H^4K} + \frac{d^{8/5}H^{11/5}K^{3/5}}{\epsilon^{2/5}}\right)$. Note that both regret bounds include two terms, in which the first term corresponds to regret of a non-private algorithm and the second term corresponds to the "cost of privacy." In both results, the non-private term matches the state-of-the-art non-private regret bound from Jin et al. (2020). However, our regret bound improves the cost of privacy from Luyo et al. (2021) in the parameters $d$ and $K$. In the most prevalent regimes of differential privacy, the privacy parameter $\varepsilon$ is chosen to be a small constant. In this case, the privacy cost term is dominated the non-private regret rate, and thus, our regret rate becomes $\widetilde{O}\left(\sqrt{d^3H^4K}\right)$, matching the non-private rate from Jin et al. (2020).

**Our techniques.** The key technical ingredient that enables our improved regret bound is *adaptivity*. In particular, our algorithm only updates its underlying policy when it detects a sufficient change in the collected data. Unlike prior work (Luyo et al., 2021) that employs an non-adaptive update schedule and triggers polynomial in $K$ number of updates, our algorithm draws the *low switching cost* techniques from Wang et al. (2021) and only triggers policy update roughly $O(\log(K))$ times. Low switching cost is particularly appealing for privacy since it largely reduce the amount of noise needed to achieve the same level of privacy parameters, which in turn leads to an improved regret.

The adaptive policy update schedule introduces challenges for the privacy analysis, since the policy update time is an unknown random variable that depends on the data and the randomness of the algorithm. In order to exploit the advantage of low switching cost, our privacy analysis needs to bound the privacy loss to be only proportional to the number updates, instead of total number of episodes $K$. To meet this challenge, our privacy analysis relies on a novel argument that models the interactions with the users as an adaptive adversary with bounded sensitivity. We believe that our low-switching-cost algorithm and analysis provide fruitful directions for answering other private RL questions in future, since they go beyond the existing paradigm that heavily rely on the binary mechanism.

### 1.2. Related Work

Our work is most related to the line of work on RL subject to JDP that includes the earlier work of Vietri et al. (2020); Garcelon et al. (2020) in the tabular settings and more recent work of (Luyo et al., 2021) in the linear MDP settings. Both (Zhou, 2022) and (Luyo et al., 2021) also consider a different form of linear function approximation called the linear mixture MDPs, which we do not study in this work. The linear mixture MDPs setting is arguably easier to learn under JDP since all the relevant sufficient statistics for learning can be written as summation and the binary mechanism is applicable. In contrast, obtaining an $O(\sqrt{K})$ regret rate in our linear MDP setting requires new techniques.

More broadly speaking, our work contributes to the growing line of work on private interactive learning, which includes the study of online learning (Guha Thakurta & Smith, 2013; Agarwal & Singh, 2017), multi-armed bandits (Mishra & Thakurta, 2015; Tossou & Dimitrakakis, 2017), linear con-

textual bandits (Shariff & Sheffet, 2018).

Technically, our algorithm is closely related to the non-private algorithms in the linear MDP settings that achives provable low regret (Jin et al., 2020). Low switching cost has also been a desideratum for RL algorithms, even absent privacy concerns. For example, (Bai et al., 2019) shows how to bound local switching cost in a tabular MDP setting: $O(H^3SA\log(K))$, and (Gao et al., 2021; Wang et al., 2021) give a provably efficient algorithm for linear MDP with low switching cost.

## 2. Preliminaries

**Notation** We use bold capital letters to denote matrices, bold lower case for vectors. Let $\boldsymbol{X}_{1:t}$ be the matrix whose rows are $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t$, then we define the Gram matrix by $\boldsymbol{\Lambda}_t = \boldsymbol{X}_{1:t}^\top \boldsymbol{X}_{1:t} = \sum_{i=1}^t \boldsymbol{x}_i \boldsymbol{x}_i^\top$. A symmetric matrix $\boldsymbol{X}$ is positive-semidefinite if $\boldsymbol{x}^\top \boldsymbol{X} \boldsymbol{x} \geq 0$ for any vector $\boldsymbol{x}$. Any such $\boldsymbol{X}$ defines a norm on vectors, so we define $\|\boldsymbol{x}\|_{\boldsymbol{X}}^2 = \boldsymbol{x}^\top \boldsymbol{X} \boldsymbol{x}$. For any positive integer $N$, we use $[N]$ to denote the set $\{1, \ldots, N\}$. We use $\boldsymbol{z} \sim \mathcal{N}(0, \sigma^2)^d$ to denote a vector-valued random variable of dimension $d$, where each coordinate is sampled i.i.d from a Gaussian distribution with variance $\sigma^2$. For a vector $\boldsymbol{x} \in \mathbb{R}^d$, we use $\|\boldsymbol{x}\|_2 = \sqrt{\sum_{i=1}^d \boldsymbol{x}_i^2}$ to denote the $\ell_2$ norm of $\boldsymbol{x}$. We can also define the spectral norm of a matrix $\boldsymbol{M}$ as $\|\boldsymbol{M}\| = \max_{\boldsymbol{x} \neq 0} \frac{\|\boldsymbol{M}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}$, which is the operator norm associated with the vector $\ell_2$ norm.

### 2.1. Markov Decision Process

We begin with the general setup for the episodic *Markov Decision Process* (MDP) before we describe the linear MDP setting we focus on. An MDP is denoted by $\text{MDP}(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$, where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, $H \in \mathbb{Z}_+$ is the length of each episode, $\mathbb{P} = \{\mathbb{P}_h\}_{h=1}^H$ are the state transition probability measures, and finally $r = \{r_h\}_{h=1}^H$ is the set of reward functions. We assume that $\mathcal{S}$ is a measurable space with possibly infinite number of elements and $\mathcal{A}$ is a finite set with cardinality $A$. For each $h \in H$, $\mathbb{P}_h(\cdot|x, a)$ denoted the transition kernel over next states if action $a$ is taken for state $x$ at step $h \in [H]$. Similarly, at each round $h$, we denote the deterministic reward function as $r_h(x_h, a_h)$ in $[0, 1]$.[1] In the first time step of each episode, the agent observes the initial state $x_1$, which can be adversarially selected. At each time-step $h \leq H$, the agent observes $x_h \in \mathcal{S}$, then picks an action $a_h \in \mathcal{A}$ and receives the reward $r_h(x_h, a_h)$. As a result, the MDP moves to a new state $x_{h+1}$ drawn from probability measures $\mathbb{P}_h(x_{h+1}|x_h, a_h)$. When state $x_{H+1}$ is reached, the episode ends and the agent receives no further reward.

[1] For notational simplicity, we study deterministic rewards, and our results can be generalized to random reward functions.

A deterministic policy $\pi$ is a function $\pi : \mathcal{S} \times [H] \to \mathcal{A}$, where $\pi(x_h, h)$ is the action that the agent takes at state $x_h$ and round $h$ in the episode. For any $h \in [H]$, we define the value function $V_h^\pi : \mathcal{S} \to \mathbb{R}$ as the expected value of cumulative rewards by following policy $\pi$ from round $h$ at an arbitrary state:

$$V_h^\pi(x) \coloneqq \mathbb{E}\left[\sum_{h'=h}^H r_{h'}(x_{h'}, \pi(x_{h'}, h'))\Big| x_h = x\right] \qquad (\forall x \in \mathcal{S}, h \in [H])$$

Moreover, we also define the action-value function $Q_h^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ to be the expected value of cumulative rewards when the agent starts from an arbitrary state-action pair at the $h$-th step and follow policy $\pi$:

$$Q_h^\pi(x, a) \coloneqq r_h(x, a) + \mathbb{E}\left[\sum_{h'=h+1}^H r_{h'}(x_{h'}, \pi(x_{h'}, h'))\Big| x_h = x, a_h = a\right] \qquad (\forall (x, a) \in \mathcal{S} \times \mathcal{A}, \forall h \in [H])$$

To simplify notation, we denote the expected value function as $[\mathbb{P}_h V_{h+1}](x, a) = \mathbb{E}_{x' \sim \mathbb{P}_h(\cdot|x,a)} V_{h+1}(x')$. The *Bellman equation* then becomes

$$\begin{aligned} Q_h^\pi(x, a) &\coloneqq \left[r_h + \mathbb{P}_h V_{h+1}^\pi\right](x, a) \\ V_h^\pi(x) &\coloneqq Q_h^\pi(x, \pi(x, h)) \\ V_{H+1}^\pi(x) &\coloneqq 0 \text{ for all } x \in \mathcal{S} \end{aligned}$$

which holds for all $(x, a) \in \mathcal{S} \times \mathcal{A}$. Similarly, the optimal Bellman equation is

$$\begin{aligned} Q_h^\star(x, a) &\coloneqq \left[r_h + \mathbb{P}_h V_{h+1}^\star\right](x, a) \\ V_h^\star(x) &\coloneqq \max_{a \in \mathcal{A}} Q_h^\star(x, a) \\ V_{H+1}^\star(x) &\coloneqq 0 \text{ for all } x \in \mathcal{S} \end{aligned}$$

The optimal policy $\pi^\star$ is a greedy policy with respect to the optimal action-value policies $\{Q_h^\star\}_{h \in [H]}$ and achieves value $\{V_h^\star\}_{h \in [H]}$.

The agent interacts with the environment over a sequence of $K$ episodes. At the beginning of each episode $k \geq 1$, the environment chooses an initial state $x_1^k$ and the agent chooses a policy $\pi_k$. Then, the agent's expected regret for the $k$-th episode is the difference in the values $V_1^\star(x_1^k) - V_1^{\pi_k}(x_1^k)$. Therefore, after playing for $K$ episodes the total expected regret is

$$R(K) = \sum_{k=1}^K \left[V_1^\star(x_1^k) - V_1^{\pi_k}(x_1^k)\right]$$

### 2.2. Linear MDP

We focus on linear *Markov Decision Process* setting, where reward and transition functions are assumed to be linear (Bradtke & Barto, 1996a; Melo & Ribeiro, 2007).

**Assumption 1** (Linear MDP). *MDP$(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$ is a linear MDP with a feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, if for any $h \in [H]$ there exist $d$ unknown measures $\mu_h = \left(\mu_h^1, \ldots, \mu_h^d\right)$ over $\mathcal{S}$ and an unknown vector $\theta_h$, such that for any $(x, a) \in \mathcal{A} \times \mathcal{S}$ we have*

$$\mathbb{P}_h(\cdot|x, a) = \langle \phi(x, a), \mu_h(\cdot) \rangle, \quad r_h(x, a) = \langle \phi(x, a), \theta_h \rangle$$

*We assume that $\|\phi(x, a)\| \leq 1$ for all $(x, a) \in \mathcal{S} \times \mathcal{A}$, and $\max\{\|\mu_h(\mathcal{S})\|, \|\theta_h\|\} \leq d$ for all $h \in [H]$.*

An important property of linear MDP is that the action-value functions are linear in the feature map $\phi$. Hence, it suffices to focus on linear action-value functions.

**Proposition 2** (Jin et al. (2020)). *For a linear MDP, for any policy $\pi$, there exist weights $\{w_h^\pi\}_{h\in[H]}$ such that for any tuple $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, we have $Q_h^\pi(x, a) = \langle \phi(x, a), w_h^\pi \rangle$*

### 2.3. (Joint) Differential Privacy in Episodic RL

In this work, we provide privacy-preserving RL algorithm that incorporates the rigorous notion of *differential privacy* (DP). We first revisit the standard DP definition. Formally, we use $U \in \mathcal{U}^K$ to denote a sequence of $K$ users participating in the RL protocol. Technically speaking, a user can be identified with a tree of depth $H$ encoding the state and reward responses they would give to all $A^H$ possible sequences of actions the agent can choose. We also say two users sequences $U$ and $U'$ are neighbors if they differ in just one user $i \leq k$.

**Definition 3** (DP (Dwork et al., 2006)). *A randomized mechanism $M$ satisfies $(\epsilon, \delta)$-differential privacy ($(\epsilon, \delta)$-DP) if for all neighboring datasets $U, U'$ that differ by one record and for all event $E$ in the output range,*

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

*When $\delta = 0$ we say pure-DP, and when $\delta > 0$, then we say approximate-DP.*

In our episodic MDP setting, there is a set of $K$ users arriving sequentially, and each user's sensitive data corresponds to the sequence of states and rewards in a single episode. We focus on the *central model* where users are willing to share their sensitive information with a trusted agent in exchange for a service or recommendation but they don't want their information to be leaked to third parties. Our goal is to avoid any inference about the user's information while interacting with the RL agent. The standard definition of DP is too stringent for this setting, since it would require the entire output, which includes the action recommendations for user in episode $k$, and cannot reveal user $k$'s private data, which prevents any utility in the recommendations. Following the prior work of Vietri et al. (2020), we consider the notion of *joint differential privacy* (JDP) (Kearns et al., 2014b), which informally requires that if any single user changes their data, the information observed by all the other $(K-1)$ users cannot not change substantially.

To introduce JDP formally, we denote $U = (u_1, \ldots, u_K)$ the user sequence that interacts with the agent over $K$ episodes. Technically speaking, a user can be identified with a tree of depth $H$ encoding the state and reward responses they would give to all the possible sequences of actions the agent can choose. In their interaction, the agent only gets to observe the information along a single root-to-leaf path in each user's tree. Let $\mathcal{U}$ denote the space of all such trees. Let $M$ be any RL algorithm that takes as input a sequence of users $U$ and outputs action recommendations over $K$ episodes. For any $k \in [K]$ we denote $M_{-k}(U)$ as all action recommendations generated by $M$ except the action during the $k$-th episode. Two user sequences $U$ and $U'$ are $k$-neighbors if they only differ in their $k$-th user.

**Definition 4.** *(Joint-Differential Privacy (JDP)) An RL algorithm $M : \mathcal{U}^K \to \mathcal{A}^{H\times[K-1]}$ is $(\epsilon, \delta)$-jointly differentially private ($(\epsilon, \delta)$-JDP) if for all $k \in [K]$, all $k$-neighboring user sequences $U, U'$, and all events $E \subseteq \mathcal{A}^{H\times[K-1]}$,*

$$\mathbb{P}\left[M_{-k}(U) \in E\right] \leq e^\epsilon \mathbb{P}\left[M_{-k}(U') \in E\right] + \delta$$

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 5** (zCDP (Bun & Steinke, 2016)). *A randomized mechanism $M$ satisfies $\rho$-Zero-Concentrated Differential Privacy ($\rho$-zCDP), if for all neighboring dataset $U, U'$ and all $\alpha \in (1, \infty)$,*

$$D_\alpha(M(U) \parallel M(U')) \leq \rho\alpha$$

*where $D_\alpha$ is the Renyi-divergence*

Any algorithm that satisfies $\rho$-zCDP also satisfies approximate-DP. The following proposition from Bun & Steinke (2016) shows how to do the mapping between zCDP and approximate-DP.

**Lemma 6.** *If $M$ satisfies $\rho$-zCDP then $M$ satisfies $\left(\rho + 2\sqrt{\rho \log(1/\delta)}, \delta\right)$-DP.*

Another basic but important property of zCDP is easy composition of zCDP mechanisms:

**Lemma 7** (zCDP Composition). *Let $M : \mathcal{U}^K \to \mathcal{Y}$ and $M' : \mathcal{U}^K \to \mathcal{Z}$ be randomized mechanisms. Suppose that $M$ satisfies $\rho$-zCDP and $M'$ satisfies $\rho'$-zCDP. Define $M'' : \mathcal{U}^K \to \mathcal{Y} \times \mathcal{Z}$ by $M''(U) = (M(U), M'(U))$. Then $M''$ satisfies $(\rho + \rho')$-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 8** ($\ell_2$-Sensitivity)**.** *Let $U \sim U'$ denote neighboring datasets. Then the $\ell_2$-sensitivity of a function $f : \mathcal{U} \to \mathbb{R}^d$ is*

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

*if $f : \mathcal{U}^K \to \mathbb{R}^{d \times d}$ is a matrix valued function, then $\Delta(f) := \max_{U \sim U'} \|f(U) - f(U')\|_{op}$.*

In our analysis, we use the Gaussian mechanism:

**Definition 9** (Gaussian mechanism)**.** *The Gaussian mechanism $\textsc{Gau}(U, f, \rho)$ takes as input a dataset $U$, a function $f : \mathcal{X} \to \mathbb{R}^d$ and a privacy parameter $\rho > 0$ and outputs $f(U) + \boldsymbol{\eta}$ where each coordinate $i \in [d]$ is sampled as $\boldsymbol{\eta}_i \sim \mathcal{N}(0, \sigma^2)$ and $\sigma^2 = \frac{\Delta(f)^2}{2\rho}$.*

We use the following facts about the Gaussian mechanism:

**Lemma 10.** *For any dataset $U \in \mathcal{U}^K$ and low sensitive function $f : \mathcal{U}^K \to \mathbb{R}^d$ and privacy parameter $\rho$. The Gaussian mechanism satisfies $\rho$-zCDP and, for $\gamma \in (0, 1)$, the error is given by:*

$$\mathbb{P}\left( \|f(U) - \textsc{Gau}(U, f, \rho)\|)_2 > \sqrt{\frac{d\Delta^2 \log(d/\gamma)}{\rho}} \right) < \gamma$$

## 3. JDP RL with Low Switching Cost

We now introduce our RL algorithm for the linear MDP setting. We will first revisit the non-private algorithm of Least-Squares Value Iterations (LSVI) and then introduce our techniques to make such an algorithm private with a desirable privacy-accuracy trade-off.

**LSVI** In the linear MDP setting, we leverage the fact that the action value function $Q_h^\star$ is linear in the feature vector (Proposition 2) and thus they can be estimated using the idea of Least-Squares Value Iterations (Bradtke & Barto, 1996b; Osband et al., 2016). In particular, the true action-value function $Q_h^\star$ is parameterized by a vector $\boldsymbol{w}_h^\star$. If we updated the policy on episode $k$, then Algorithm 1 recursively estimates $\boldsymbol{w}_h^\star$ for all $h$ using LSVI. Let $Q_h^k$ and $\boldsymbol{w}_h^k$ denote the running estimates of the action-value functions and their corresponding parameters over episodes. In each episode $k$, the algorithm computes the following sufficient statistics relevant to the estimation of $\boldsymbol{w}_h^\star$: For some $\lambda > 0$, starting with $h = H$ down to $h = 1$, compute a least-squares estimate $\boldsymbol{w}_h^k$ via

$$\boldsymbol{\Lambda}_h^k = \lambda \boldsymbol{I} + \sum_{i=1}^{k-1} \phi\left(x_h^i, a_h^i\right) \phi\left(x_h^i, a_h^i\right)^\top \tag{1}$$

$$\boldsymbol{y}_h^k = \sum_{i=1}^{k-1} \phi\left(x_h^i, a_h^i\right) \left(r_h^i + V_{h+1}^k(x_{h+1}^i)\right) \tag{2}$$

$$\boldsymbol{w}_h^k = \left(\boldsymbol{\Lambda}_h^k\right)^{-1} \boldsymbol{y}_h^k, \tag{3}$$

which subsequently define the the corresponding optimistic estimates for action-value and state-value functions:

$$Q_h^k(x, a) = \langle \phi(x, a), \boldsymbol{w}_h^k \rangle + \beta \|\phi(\cdot, \cdot)\|_{\left(\boldsymbol{\Lambda}_h^k\right)^{-1}} \tag{4}$$

$$V_h^k(x) = \max_a Q_h^k(x, a) \tag{5}$$

Given these optimistic estimates, the non-private algorithm follows a greedy policy with respect to the estimates of action-value functions.

**Private LSVI-UCB** To provide a JDP variant of the LSVI algorithm, our strategy is to identify a collection of statistics that are compatible with differentially private releases and sufficient for constructing estimates for the action value functions.

**Binary Mechanism** We first introduce the mechanism to privatize the sequence of Gram matrices $\boldsymbol{\Lambda}_h^k$, defined in (1). First, notice that the Gram matrix statistic is in form of a sum. This allows us to use a variation of the binary mechanism due to Shariff & Sheffet (2018), which is a tree-based aggregation mechanism that sequentially releases sums of matrices privately. Our goal is to use the binary mechanism to produces a sequence of privatized gram matrices, which we denote by $\{\widetilde{\boldsymbol{\Lambda}}_h^k\}_{k,h \in [K] \times [H]}$.

**Gaussian mechanism with low switching cost** Our second set of statistics is $\boldsymbol{y}_h^k$ (defined in Equation (2)). The challenge with privatizing $\boldsymbol{y}_h^k$ is that it does not take the form of summation, as it depends on the value function estimate $V_{h+1}^k$ for episode $k$. As a result, we cannot apply the binary mechanism. Therefore, we use the Gaussian mechanism on each update episode to privatize $\widetilde{\boldsymbol{y}}_h^k$. However, adding Gaussian noise to every episode would lead to too much noise. For this reason, we minimize the number of episodes we update the policy.

In Algorithm 1, an update episode occurs every time the determinant of the noisy gram matrix $\widetilde{\boldsymbol{\Lambda}}_h^k$ doubles for any $h \in [H]$. Since Algorithm 1 does not update the action-value function every round, we use $\tilde{k}$ to denote the last episode the policy $\pi_h^k$ was updated.

As shown in (Wang et al., 2021), this adaptive update schedule allows us to bound the number of policy updates by

$O(dH\log(K))$ while increasing the regret only by a factor of 2. Therefore, instead of applying composition over $K$ Gaussian mechanisms, we only have $O(dH\log(K))$. Having $O(\log(K))$ updates is critical for obtaining our improved regret rate. The algorithm from (Luyo et al., 2021) uses a non-adaptive batching technique that leads to a number of updates polynomial in $K$. In the next section, we show how to analyze the privacy guarantee of this adaptive update algorithm.

**Algorithm 1** Private LSVI-UCB

**input** Privacy parameter $\rho$, policy update rate $C$, fail probability $p$, confidence width $\beta$.

Notation: Let $\phi(x_h^i, a_h^i) = \phi_h^i$

$\widetilde{\lambda}_\Lambda := \frac{\log(K)\left(6\sqrt{d+1}+2\log\left(\frac{3KH}{p}\right)\right)}{\sqrt{2\rho}}$

$N_{\max} := \frac{dH}{\log 2}\log\left(1+\frac{K}{\widetilde{\lambda}_\Lambda d}\right)$

**Initialize:** For all $h\in[H]: \widetilde{\mathbf{\Lambda}}_h^1 \leftarrow 2\widetilde{\lambda}_\Lambda \boldsymbol{I}$,

**Initialize:** A counter $\mathcal{C}_h^{(\cdot)}(\frac{\rho}{2H})$ as in Equation (6).

**Initialize:** $\tilde{k}\leftarrow 1$, and $N_{\text{count}}\leftarrow 1$.

**for** $k=1,\dots,K$ **do**

  **if** $\exists h\in[H], \frac{\det\left(\widetilde{\mathbf{\Lambda}}_h^k\right)}{\det\left(\widetilde{\mathbf{\Lambda}}_h^{\tilde{k}}\right)} \geq C$ **and** $N_{\text{count}} < N_{\max}$ **then**

    **for** $h=H,\cdots,1$ **do**

      $\boldsymbol{y}_h^k \leftarrow \sum_{i=1}^{k-1}\phi_h^i(r_h^i + V_{h+1}^k(x_{h+1}^i))$

      $\widetilde{\boldsymbol{y}}_h^k \sim \mathcal{N}\left(\boldsymbol{y}_h^k, \frac{6H^2N_{\max}}{\rho}\boldsymbol{I}\right)^d$

      $\boldsymbol{w}_h^k \leftarrow \left(\widetilde{\mathbf{\Lambda}}_h^k\right)^{-1}\widetilde{\boldsymbol{y}}_h^k$

      $\mathrm{B}_h^k(\cdot,\cdot) = \beta\left[\phi(\cdot,\cdot)^\top(\widetilde{\mathbf{\Lambda}}_h^k)^{-1}\phi(\cdot,\cdot)\right]^{\frac{1}{2}}$

      $Q_h^k(\cdot,\cdot) = \min\{\phi(\cdot,\cdot)^\top\boldsymbol{w}_h^k + \mathrm{B}_h^k(\cdot,\cdot), H\}$

      Update $\tilde{k}\leftarrow k$ **and** $N_{\text{count}}\leftarrow N_{\text{count}}+1$

    **end for**

  **end if**

  **for** step $h=1\dots,H$ **do**

    Observe $x_h^k$

    Take action $a_h^k \leftarrow \arg\max_{a\in\mathcal{A}} Q_h^{\tilde{k}}(x_h^k, a)$.

    Observe reward $r_h^k \leftarrow r_h(x_h^k, a_h^k)$ .

    Update $\mathcal{C}_h^{(\cdot)}(\frac{\rho}{2H})$ with $\phi_h^k(\phi_h^k)^\top$ and set

$$\widetilde{\mathbf{\Lambda}}_h^{k+1} \leftarrow \mathcal{C}_h^{(k)}(\tfrac{\rho}{2H}) + 2\widetilde{\lambda}_\Lambda \boldsymbol{I}$$

  **end for**

**end for**

## 4. Analysis

We begin by analyzing the switching cost of Algorithm 1, then the privacy and regret guarantees.

### 4.1. Switching Cost

We begin by analysing the switching cost of Algorithm 1, which will be used in Section 4.2 for the privacy analysis .

**Theorem 11.** *[Similar to Lemma C.3 in (Wang et al., 2021)] Let $\widetilde{\lambda}_\Lambda$ be defined as in Algorithm 1. Condition on the event that $\left\|\widetilde{\mathbf{\Lambda}}_h^k - \mathbf{\Lambda}_h^k\right\|_{op} \leq \widetilde{\lambda}_\Lambda$ for all $h,k\in[H]\times[K]$. For $C=2$ and $\widetilde{\lambda}_\Lambda > 0$, the global switching cost of Algorithm 1 is bounded by: $N_{count} \leq \frac{dH}{\log 2}\log\left(1+\frac{K}{\widetilde{\lambda}_\Lambda d}\right)$.*

*Proof Sketch.* See the full proof derivation in Appendix C. We use a determinant-based analysis, similar to that of the non-private algorithm in (Wang et al., 2021). The main difference in our proof is that the determinant of $\widetilde{\mathbf{\Lambda}}_h^{\tilde{k}}$ is now at most $(\widetilde{\lambda}_\Lambda + (k-1)/d)^d$ to account for the perturbation from binary mechanism. Also, we notice that every time the update criteria is met, the determinant of $\widetilde{\mathbf{\Lambda}}_h^{k+1}$ increases by at most twice. Hence, in total we can bound the number of updates by $O\left(\frac{dH}{\log 2}\log(1+\frac{K}{\widetilde{\lambda}_\Lambda d})\right)$. □

The low switching cost guarantee above is crucial for our algorithm to achieve our improved regret guarantee since it allows us to reduce the level of noise to preserve privacy.

### 4.2. Privacy Analysis

This section provides the privacy analysis of Algorithm 1. But, first, we state our main privacy guarantee:

**Theorem 12** (Privacy)**.** *Algorithm 1 satisfies $(\epsilon,\delta)$-JDP.*

For the privacy analysis of Algorithm 1, we use zCDP (Bun & Steinke, 2016) because it simplifies the composition of Gaussian mechanisms. At the end, we translate our results in terms of approximate DP. In this section, we show that releasing statistics $\{\widetilde{\mathbf{\Lambda}}_h^k\}_{k,h\in[K]\times[H]}$ satisfies $(\rho/2)$-zCDP and then show that releasing $\{\widetilde{\boldsymbol{y}}_h^{\tilde{k}}\}_{k,h\in[K]\times[H]}$ also satisfies $(\rho/2)$-zCDP. Thus, by composition of zCDP mechanisms (Theorem 7), we have that releasing the sufficient statistics satisfies $\rho$-zCDP and, by Theorem 6, also satisfies $(\epsilon,\delta)$-DP for $\epsilon = \rho + \sqrt{\rho\log(1/\delta)}$.

It follows that Algorithm 1 is $(\epsilon,\delta)$-JDP by the Billboard lemma due to Hsu et al. (2016), since the actions sent to each user depends on a function constructed with DP and their private data only. We state the billboard lemma here:

**Lemma 13** (Billboard lemma (Hsu et al., 2016))**.** *Suppose that randomized mechanism $M:\mathcal{U}\to\mathcal{R}$ is $(\epsilon,\delta)$-differentially private. Let $U\in\mathcal{U}$ be a dataset containing $n$ users. Then, consider any set of functions $f_i: \mathcal{U}_i\times\mathcal{R}\to\mathcal{R}_i$, for $i\in[n]$, where $\mathcal{U}_i$ is the portion of the database containing the $i$'s user data. Then the com-*

*position* $\{f_i(\mathrm{proj}_i(U), M(U))\}_{i\in[n]}$ *is* $(\epsilon,\delta)$*-JDP, where* $\mathrm{proj}: \mathcal{U} \to \mathcal{U}_i$ *is the projection to* $i$*'s data.*

We start with bounding the sensitivity of the two types of sufficient statistics.

**Lemma 14** ($\ell_2$-sensitivity)**.** *For any* $(k,h) \in [K]\times[H]$*. We have* $\Delta\left(\boldsymbol{\Lambda}_h^k\right) \le 2$ *and* $\Delta\left(\boldsymbol{y}_h^k\right) \le 2H+2$*.*

*Proof Sketch.* In order to bound the sensitivity of $\boldsymbol{y}_h^k$, we need to consider any two neighboring user sequences $U$ and $U'$ with outputs $\boldsymbol{y}_h^k$ and $(\boldsymbol{y}')_h^k$, respectively. Let $i \le k$ be some episode with $x_h^i \neq {x'}_h^i$ and $a_h^i \neq {a'}_h^i$, where $(x_h^i, a_h^i) \in U$ and $({x'}_h^i, {a'}_h^i) \in U'$. By definition of neighboring user sequences and triangle inequality, the difference between $\boldsymbol{y}_h^k$ and $(\boldsymbol{y}')_h^k$ can be written as the sum of $\left\|\phi(x_h^i,a_h^i)\left(r_h(x_h^i,a_h^i)+V_{h+1}^k(x_{h+1}^i)\right)\right\|_2$ and $\left\|\phi({x'}_h^i,{a'}_h^i)\left(r_h({x'}_h^i,{a'}_h^i)+V_{h+1}^i({x'}_{h+1}^i)\right)\right\|_2$. Also, by assumption of linear MDP, we have $\|\phi(\cdot,\cdot)\| \le 1$ and $r_h(\cdot,\cdot) \le 1$, so the first summand in both terms is at most 1. Furthermore, $V_{h+1}^i(x_{h+1}^i)$ and $V_{h+1}^i({x'}_{h+1}^i)$ are at most $H$. Using the assumption that $\|\phi(\cdot,\cdot)\| \le 1$, the second summand in each term is at most $H$. Combining these bounds, we arrive at the claim.

By definition of neighboring sequences and triangle inequality, the sensitivity of $\boldsymbol{\Lambda}_h^k$ can be written as the sum of $\left\|\phi(x_h^k,a_h^k)\phi(x_h^k,a_h^k)^\top\right\|$ and $\left\|\phi({x'}_h^k,{a'}_h^k)\phi({x'}_h^k,{a'}_h^k)^\top\right\|$. Since $\|\phi(\cdot,\cdot)\| \le 1$, the sensitivity of $\boldsymbol{\Lambda}_h^k$ is at most 2. □

We release these two types of statistics with two mechanisms. Next, we show that both the Binary mechanism and the Gaussian mechanism each satisfies $\rho/2$-zCDP.

**Analysis of Binary Mechanism:** First, we use the Gaussian binary mechanism (Shariff & Sheffet, 2018) (a variant of the binary mechanism due to (Chan et al., 2011; Dwork et al., 2010) that preserves positive definiteness (PD) in matrices) to privatize the statistics $\{\boldsymbol{\Lambda}_h^k\}_{k,h\in[K]\times[H]}$. Here we show that releasing the sequence of private Gram matrices, denoted by $\{\widetilde{\boldsymbol{\Lambda}}_h^k\}_{k,h\in[K]\times[H]}$, satisfies $\rho/2$-zCDP and also, with high probability, the all are PD matrices.

The algorithm initializes $H$ private counters $\mathcal{C}_h^{(\cdot)}(\frac{\rho}{2H})$ for $h\in[H]$ such that each satisfies $(\rho/(2H))$-zCDP. Denote $\phi_h^k = \phi\left(x_h^k, a_h^k\right)$ as the data observed during episode $k$. Each counter $\mathcal{C}_h^{(\cdot)}(\frac{\rho}{2H})$ observes a stream of Gram matrices $U_h = (\phi_h^1{\phi_h^1}^\top, \ldots, \phi_h^K{\phi_h^K}^\top)$ over $K$ episodes and maintains a binary tree in which each internal node represents a partial sum. Let $\Sigma^{i:j}$ be the function of the input stream $U_h$ that computes the partial sum of all events between between time $i$ and $j$, defined as $\Sigma^{i:j}(U_h) = \sum_{\tau=i}^{j-1}\phi_h^\tau(\phi_h^\tau)^\top$. The sum of all events before episode $k$ is can be computed by a function $\Sigma^{1:k}$, which is the output of at most $\log(K)$ partial sums. Note that $\Delta(\Sigma^{1:k}) \le \log(K)$. On each episode $k$, the private counter for $h$ outputs a random matrix $\mathcal{C}_h^{(k)}(\frac{\rho}{2H})$ using the Gaussian mechanism as follows:

$$\mathcal{C}_h^{(k)}(\tfrac{\rho}{2H}) \leftarrow \textsc{Gau}\left(U_h, \Sigma^{1:k}, \frac{\log(K)^2}{2\rho}\right) \tag{6}$$

Each partial sum satisfies $\left(\frac{\rho}{2H\log(K)}\right)$-zCDP. However, since any single data point in $U_h$ appears in at most $\log(K)$ nodes of $\mathcal{C}_h^{(\cdot)}(\frac{\rho}{2H})$, we must do composition over $\log(K)$ events. It follows that the output of $\mathcal{C}_h^{(\cdot)}(\frac{\rho}{2H})$ satisfies $\frac{\rho}{2H}$-zCDP. By another round of composition over all time steps $H$, we have that releasing $K\cdot H$ counts given by $\{\mathcal{C}_h^{(k)}(\frac{\rho}{2H})\}_{k\in[K],h\in[H]}$ satisfies $\rho/2$-zCDP.

The noise added by the BinMechcan violate the requirement that $\widetilde{\boldsymbol{\Lambda}}_h^k$ is positive definite (PD). For that reason, in Equation (6), we shift the noisy Gram matrix of $\mathcal{C}_h^{(\cdot)}(\frac{\rho}{2H})$ by a constant $2\widetilde{\lambda}_\Lambda \boldsymbol{I}$.

Next we show that the shifted noisy Gram matrix $\widetilde{\boldsymbol{\Lambda}}_h^k$ is PD, by showing that all eigenvalues are strictly positive. First, let $\boldsymbol{M}_h^k$ be the noise added by the BinMech on episode $k$ such that $\widetilde{\boldsymbol{\Lambda}}_h^k = \sum_{\tau=1}^{k-1}\phi_h^\tau(\phi_h^\tau)^\top + 2\widetilde{\lambda}_\Lambda\boldsymbol{I} + \boldsymbol{M}_h^k$, then we only need to show that $2\widetilde{\lambda}_\Lambda\boldsymbol{I} + \boldsymbol{M}_h^k$ is PD. By known concentration bounds (Tao, 2012) on the matrix operator norm state that with probability at least $1-p$ on all for $k,h\in[K]\times[H]$:

$$\|\boldsymbol{M}_h^k\|_{op} \le \widetilde{\lambda}_\Lambda := \frac{\log(K)\left(6\sqrt{d+1}+2\log\left(\frac{3KH}{p}\right)\right)}{\sqrt{2\rho}} \tag{7}$$

If the eigenvalues of $\boldsymbol{M}_h^k$ are $\lambda_1,\ldots,\lambda_d$, then $\boldsymbol{M}_h^k + 2\widetilde{\lambda}_\Lambda\boldsymbol{I}$ has eigenvalues $\lambda_1+2\widetilde{\lambda}_\Lambda,\ldots,\lambda_d+2\widetilde{\lambda}_\Lambda$. By definition of operator norm and Equation (7) we have $\max_{i\in[d]}|\lambda_i| = \|\boldsymbol{M}_h^k\|_{op} \le \widetilde{\lambda}_\Lambda$. Therefore, all eigenvalues of $\widetilde{\boldsymbol{\Lambda}}_h^k$ are positive and thus the matrix is PD.

**Analysis of Gaussian Mechanism:** For the remaining of this section, the focus is to show that releasing the statistics $\{\widetilde{\boldsymbol{y}}_h^k\}_{k,h\in[K]\times[H]}$ also satisfies $(\rho/2)$-zCDP. Recall that, if $k$ is an update episode then Algorithm 1 adds Gaussian noise to the statistics $\boldsymbol{y}_h^k$. Naively, we could update the episode every round, but the noise added would scale with $O(\rho/K)$, giving an error in the order of $O(\sqrt{K/(\rho)})$, which leads to sub-optimal regret on the variable $K$. Therefore, one would want to update the policy at most $O(dH\log(K))$ times because it allows us to scale the noise with $\rho/\log(K)$, which gives logarithmic error on the parameter $K$.

The main challenge is that the update episode is a random variable dependent on the user sequence. Thus, any single user in the sequence affects the future update episodes.

Therefore, to simplify the analysis, we pretend to run a hypothetical Gaussian mechanism to release a constant function during rounds in which an update is not triggered. Formally, for any $h \in [H]$, we construct an adversary that selects an adaptive sequence of functions $f_h^1, \ldots, f_h^K$ as follows: On episode $k$, if $k$ is an update episode then the adversary sets $f_h^k(U_k) = \sum_{i=1}^{k-1} \phi\left(x_h^i, a_h^i\right) r_h^i := \boldsymbol{y}_h^k$ for all $h \in [H]$. Otherwise, the adversary sets $f_h^k(U_k) = 0$ for all $h \in [H]$. On non-update episodes, $f_h^k$ is a constant function independent of the data. Also, Algorithm 1 uses $f_h^k$ only on update episodes to modify the policy and ignores it otherwise.

The following lemma gives a tool to analyze the protocol under zCDP as long as the joint sensitivity of the sequences $\{f_h^1, \ldots, f_h^K\}_{h\in[K]}$ is bounded.

**Lemma 15.** *For any $K > 0$, let $\{f_k : \mathcal{U} \to \mathbb{R}^d\}_{k\in[K]}$ be a sequence of adaptively chosen functions, such that the joint sensitivity is bounded by: $\sum_{k=1}^K \Delta(f_i) \le \Delta$. Then, the composition of $\{f_k(D_k^b) + \eta_k\}_{k\in[K]}$, where $\eta_k \sim \mathcal{N}(0, \frac{\Delta}{\rho})$ satisfies $\rho$-zCDP.*

It only remains to show that the joint sensitivity of the adaptive sequences is bounded. First, we show that each $f_h^k$ has bounded sensitivity. By definition of $f_h^k$ and Theorem 14, we have on any update episode its sensitivity is given by $\Delta(f_h^k) = \Delta(\boldsymbol{y}_h^k) \le 3H$ , and on non-update episodes we have that the sensitivity is $\Delta(f_h^k) = 0$ for any $h \in [H]$.

Next, we use the fact that number of update episodes is bounded by $N_{\max} := \frac{dH}{\log 2} \log\left(1 + \frac{K}{\lambda_\Lambda d}\right)$. Then the joint sensitivity of the the adaptive sequences is given by

$$\sum_{h=1}^{H}\sum_{k=1}^{K} \Delta(f_h^k) = \sum_{h=1}^{H}\sum_{k=1}^{K} \Delta(\boldsymbol{y}_h^k) \le 3H^2 N_{\max}$$

Therefore releasing $f_h^k$ with a $\left((1/2)\rho/\left(3H^2 N_{\max}\right)\right)$-zCDP Gaussian mechanism on every episode $k \in [K]$ satisfies $\rho/2$-zCDP. This concludes the proof.

### 4.3. Regret Analysis

After showing that Algorithm 1 is JDP, we finish the analysis by showing the regret guarantees in the following theorem.

**Theorem 16** (Regret)**.** *For any $p \in (0, 1)$, any privacy parameter $\rho > 0$, if we set $C = 2$ and*

$$\beta = 5H^2\sqrt{d\widetilde{\lambda}_\Lambda \log(\chi)} + 6dH\sqrt{\log(\chi)}$$

*with $\chi = \frac{25^4 \cdot 162 \cdot K^4 \cdot d \cdot U_K \cdot H}{p}$ in Algorithm 1, then with probability $1 - p$ , the total regret of Private LSVI-UCB (Algorithm 1) is at most Algorithm 1 satisfies $(\epsilon, \delta)$-differential privacy with $\epsilon = \rho + 2\sqrt{\rho \log(1/\delta)}$. If $\epsilon < 1$, then the total regret of Algorithm 1 is at most*

$$R(K) \le \widetilde{O}\left(d^{3/2}H^2K^{1/2} + \frac{H^3 d^{5/4} K^{1/2}\left(\log\left(\frac{1}{\delta}\right)\right)^2}{\epsilon^{1/2}}\right)$$

*Proof Sketch.* We present a proof sketch here for the regret analysis. Here the $\widetilde{O}(\cdot)$ notation hides all log and constants terms. See Appendix B for full detail of the proof. The analysis is similar to Jin et al. (2020), but there are several challenges we need to handle due to privacy noise. The proof structure is as follows: (1) Show that the privacy cost to preserve the set of sufficient statistics scales with $O(\log(K)$. (2) Construct new confidence bounds given the private statistics and show that the action-value function from Algorithm 1 is optimistic. (3) As in Jin et al. (2020) and Wang et al. (2021), we follow an optimism strategy and upper bound the regret by the sum of bonus terms: $\sum_{k=1}^k \sum_{h=1}^H \mathrm{B}_h^k(x_h^k, a_h^k)$, where $B$ is defined in line 13 of Algorithm 1.

**Private Least Squares** As described in Section 3, Algorithm 1 uses a Least-Squares-Value-Iteration (LSVI) to estimate the action-value function $Q_h^\star := \langle \phi(x, a), \boldsymbol{w}_h^\star \rangle$, but to preserve privacy, Algorithm 1 adds noise to the sufficient statistics. For any $k, h \in [K] \times [H]$, let $\boldsymbol{\Lambda}_h^k$ be the gram matrix as defined in Equation (1) and $\boldsymbol{y}_h^k$ be as defined in Equation (2). Let their privatized statistics be

$$\widetilde{\boldsymbol{\Lambda}}_h^k = \boldsymbol{\Lambda}_h^k + 2\widetilde{\lambda}_\Lambda \boldsymbol{I} + \boldsymbol{M}_h^k, \qquad \widetilde{\boldsymbol{y}}_h^k = \boldsymbol{y}_h^k + \boldsymbol{\eta}_h^k \tag{8}$$

where $\{\boldsymbol{M}_h^k, \boldsymbol{\eta}_h^k\}_{k,h\in[K]\times[H]}$ are noise injected to privatize the statistics. Let $\lambda_{\min}(\cdot)$ denote the minimum eigenvalue of a matrix, then by Equation (7) we have a lower bound on the minimum eigenvalue of $\widetilde{\boldsymbol{\Lambda}}_h^k$:

$$\lambda_{\min}(\widetilde{\boldsymbol{\Lambda}}_h^k) \ge \widetilde{\lambda}_\Lambda := \frac{\log(K)\left(6\sqrt{d+1} + 2\log\left(\frac{3KH}{p}\right)\right)}{\sqrt{2\rho}} \tag{9}$$

Also, by the tail bound on the Gaussian noise we have $\|\boldsymbol{\eta}_h^k\|_2 \le \widetilde{O}\left(Hd/\sqrt{\rho}\right)$. It follows that

$$\|\boldsymbol{\eta}_h^k\|_{(\widetilde{\boldsymbol{\Lambda}}_h^k)^{-1}} \le \frac{1}{\sqrt{\widetilde{\lambda}_\Lambda}}\|\boldsymbol{\eta}_h^k\|_2 \le \widetilde{O}\left(H^2\sqrt{d\widetilde{\lambda}_\Lambda}\right) \tag{10}$$

Recall that LSVI, on episode $k$, starts with $h = H$ and recursively computes $\boldsymbol{w}_h^k = (\widetilde{\boldsymbol{\Lambda}}_h^k)^{-1}\widetilde{\boldsymbol{y}}_h^k$.

**Upper Confidence Bounds (UCB)** Let $\pi$ be any policy with corresponding action-value $Q_h^\pi(x, a) = \langle \phi(x, a), \boldsymbol{w}_h^\pi \rangle$. Let $\widehat{Q}_h^k(x, a) = \left\langle \phi(x, a), \boldsymbol{w}_h^k \right\rangle$ be the private empirical action-value function induced by $\boldsymbol{w}_h^k$ without the bonus term. We give a confidence bound around $\widehat{Q}_h^k$ and construct an optimistic action-value function $Q_h^k$ (i.e., $Q_h^k(x, a) \ge$

$Q_h^\star(x,a)$ for all $(x,a)$), taking into account the noise added to privatize the statistics. We begin by decomposing the term $\boldsymbol{w}_h^k - \boldsymbol{w}_h^\pi$ as follows:

$$\boldsymbol{w}_h^k - \boldsymbol{w}_h^\pi = \underbrace{\left(\widetilde{\boldsymbol{\Lambda}}_h^k\right)^{-1}\left(\boldsymbol{y}_h^k\right) - \boldsymbol{w}_h^\pi}_{\boldsymbol{a}} + \underbrace{\left(\widetilde{\boldsymbol{\Lambda}}_h^k\right)^{-1}\boldsymbol{\eta}_h^k}_{\boldsymbol{b}}$$

Then, for all $(x,a,h,k)$, we can write

$$\widehat{Q}_h^k(x,a) - Q_h^\pi(x,a) = \langle \phi(x,a), \boldsymbol{a}+\boldsymbol{b}\rangle$$

and bound each term independently. Next we directly use the Lemma B.4 in Jin et al. (2020) to bound $\langle \phi(x,a),\boldsymbol{a}\rangle$ by the expected difference at next step, plus an error term that depends on the minimum eigenvalue of the gram matrix, which in our case is $\widetilde{\lambda}_\Lambda$. Therefore, we obtain:

$$|\langle \phi(x,a),\boldsymbol{a}\rangle| \le \mathbb{P}_h(V_h^k - V_h^\pi)(x,a) + \widetilde{O}\left(H\sqrt{d\widetilde{\lambda}_\Lambda} + dH\right)\|\phi(x,a)\|_{(\widetilde{\boldsymbol{\Lambda}}_h^k)^{-1}} \quad (11)$$

where $\mathbb{P}_h(V_h^k - V_h^\pi)(x,a)$ is the expected error in time step $h+1$ between $V_{h+1}^k$ and $V_{h+1}^\pi$ after taking action $(x,a)$. It remains to give a confidence bound for the second term $\langle \phi(x,a),\boldsymbol{b}\rangle$, which follows simply by the Cauchy-Schwartz inequality:

$$|\langle \phi(x,a),\boldsymbol{b}\rangle| \le \left\|\boldsymbol{\eta}_h^k\right\|_{(\widetilde{\boldsymbol{\Lambda}}_h^k)^{-1}} \cdot \|\phi(x,a)\|_{(\widetilde{\boldsymbol{\Lambda}}_h^k)^{-1}} \quad (12)$$

Note that by the bound in (10), the term $\left\|\boldsymbol{\eta}_h^k\right\|_{(\widetilde{\boldsymbol{\Lambda}}_h^k)^{-1}}$ is smaller than the second term in (11). Therefore, choosing $\beta = \widetilde{O}\left(H^2\sqrt{d\widetilde{\lambda}_\Lambda} + dH\right)$ is enough to upper bound both (11) and (12). That is, with high probability, for any policy $\pi$ and all $(x,a,k,h)$ we have:

$$\widehat{Q}_h^k(x,a) - Q_h^\pi(x,a) \le \mathbb{P}_h(V_h^k - V_h^\pi)(x,a) + \beta\|\phi(x,a)\|_{(\widetilde{\boldsymbol{\Lambda}}_h^k)^{-1}} \quad (13)$$

Furthermore, by the upper bound (13) the optimistic action-value function is given by $Q_h^k(x,a) = \left\langle \phi(x,a), \boldsymbol{w}_h^k\right\rangle + \beta\|\phi(x,a)\|_{(\widetilde{\boldsymbol{\Lambda}}_h^k)^{-1}}$. That is for all $(x,a,k,h)$:

$$Q_h^k(x,a) \ge Q_h^\star(x,a) \quad (14)$$

**Regret proof** In this section we use the following notation $\phi_h^k = \phi\left(x_h^k, a_j^k\right)$. Note that on episode $k$, the agent acts according to policy $\pi_k$ which is given by $\pi_k(x,h) = \arg\max_a Q_h^{\tilde{k}}(x,a)$, where $\tilde{k} \le k$ be the last episode the algorithm updated its policy for time step $h$.

By optimism (14), the regret can be upper bounded by the difference of the optimistic value function $V_h^k(\cdot) = \max_a Q_h^k(\cdot,a)$ and the value function induced by policy $\pi_k$:

$$R(K) = \sum_{k=1}^K V_1^\star(x_1^k) - V_1^{\pi_k}(x_1^k) \le \sum_{k=1}^K V_1^{\tilde{k}}(x_1^k) - V_1^{\pi_k}(x_1^k)$$

For the next step we set $\delta_h^k = V_h^{\tilde{k}}(x_h^k) - V_h^{\pi_k}(x_h^k)$ and $\zeta_{h+1}^k = \mathbb{P}_h(V_h^{\tilde{k}} - V_h^\pi)(x_h^k,a_h^k) - \delta_{h+1}^k$. Then by the bound in (13), we can recursively relate the error on time step $h$ to the error on time step $h+1$. The results is:

$$\delta_h^k \le \delta_{h+1}^k + \underbrace{\mathbb{P}_h(V_h^{\tilde{k}} - V_h^\pi)(x_h^k,a_h^k) - \delta_{h+1}^k}_{\zeta_{h+1}^k} + \beta\|\phi_h^k\|_{(\widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}})^{-1}}$$

Unfolding the recursion, we have that with high probability, the regret of Algorithm 1 is upper bounded by:

$$R(K) \le \underbrace{\sum_{k=1}^K\sum_{h=1}^H \tilde{\zeta}_h^k}_{\le\widetilde{O}\left(\sqrt{KH^3}\right)} + \underbrace{3\beta\sum_{k=1}^K\sum_{h=1}^H \beta\|\phi_h^k\|_{(\widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}})^{-1}}}_{\le\widetilde{O}\left(H\beta\sqrt{dK}\right)} \quad (15)$$

The first term of (15) is the sum of a zero-mean martingale difference sequence bounded by $2H$. Thus, using standard concentration inequalities it's at most $\widetilde{O}\left(2\sqrt{KH^3}\right)$ with high probability. For the second term, the first step is to show that the low switching cost constraint only adds a constant to the regret. This can be seen by the following inequality: The $\|\phi_h^k\|_{(\widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}})^{-1}} \le \sqrt{2}\|\phi_h^k\|_{(\widetilde{\boldsymbol{\Lambda}}_h^k)^{-1}}$. The final bound of the second term in (15) follows from an application of Cauchy-Schwartz inequality and the trace-determinant lemma and elliptical potential lemma from Abbasi-Yadkori et al. (2011). The final bound on the second term is $\widetilde{O}\left(H\beta\sqrt{dK}\right)$. □

## 5. Conclusion

Our algorithm follows an adaptive policy update schedule and benefits from low switching cost, i.e only performs $O(\log(K))$ updates, compared to $O(\text{poly}(K))$ updates in previous works using non-adaptive batching. However, the update time is an unknown random variable depending on the data and randomness of the algorithm, which introduces new challenges in privacy analysis. We view the interactions between the algorithm and the users as an adaptive adversary with bounded sensitivity. Hence, our privacy cost is only proportional to the number of updates, instead of the number of episodes. Our low-switching cost algorithm and analysis advances the existing paradigm using binary mechanism and provides a foundation for future works in private RL.

## Acknowledgement

ZSW, DN, GV were supported in part by the NSF FAI Award #1939606, NSF SCC Award #1952085, a Google Faculty Research Award, a J.P. Morgan Faculty Award, a Facebook Research Award, and a Mozilla Research Grant. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and not necessarily reflect the views of the National Science Foundation and other funding agencies.

## References


Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24:2312–2320, 2011.

Agarwal, N. and Singh, K. The price of differential privacy for online learning. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 32–40. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/agarwal17a.html.

Bai, Y., Xie, T., Jiang, N., and Wang, Y.-X. Provably efficient q-learning with low switching cost. In *Advances in Neural Information Processing Systems*, pp. 8004–8013, 2019.

Bradtke, S. J. and Barto, A. G. Linear least-squares algorithms for temporal difference learning. *Mach. Learn.*, 22(1–3):33–57, jan 1996a. ISSN 0885-6125. doi: 10.1007/BF00114723. URL https://doi.org/10.1007/BF00114723.

Bradtke, S. J. and Barto, A. G. Linear least-squares algorithms for temporal difference learning. *Machine learning*, 22(1):33–57, 1996b.

Bun, M. and Steinke, T. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pp. 635–658. Springer, 2016.

Chan, T.-H. H., Shi, E., and Song, D. Private and continual release of statistics. *ACM Transactions on Information and System Security (TISSEC)*, 14(3):1–24, 2011.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006.

Dwork, C., Naor, M., Pitassi, T., and Rothblum, G. N. Differential privacy under continual observation. In *Proceedings of the Forty-Second ACM Symposium on Theory of Computing*, STOC '10, pp. 715—-724, New York, NY, USA, 2010. Association for Computing Machinery.

Gao, M., Xie, T., Du, S. S., and Yang, L. F. A provably efficient algorithm for linear markov decision process with low switching cost. *arXiv preprint arXiv:2101.00494*, 2021.

Garcelon, E., Perchet, V., Pike-Burke, C., and Pirotta, M. Local differentially private regret minimization in reinforcement learning. *arXiv preprint arXiv:2010.07778*, 2020.

Guha Thakurta, A. and Smith, A. (nearly) optimal algorithms for private online learning in full-information and bandit settings. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/file/c850371fda6892fbfd1c5a5b457e5777-Paper.pdf.

Hsu, J., Huang, Z., Roth, A., Roughgarden, T., and Wu, Z. S. Private matchings and allocations. *SIAM Journal on Computing*, 45(6):1953–1984, 2016.

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

Kearns, M., Pai, M., Roth, A., and Ullman, J. Mechanism design in large games: Incentives and privacy. In *Proceedings of the 5th Conference on Innovations in Theoretical Computer Science*, ITCS '14, pp. 403–410, New York, NY, USA, 2014a. Association for Computing Machinery. ISBN 9781450326988. doi: 10.1145/2554797.2554834. URL https://doi.org/10.1145/2554797.2554834.

Kearns, M., Pai, M. M., Roth, A., and Ullman, J. Mechanism design in large games: Incentives and privacy. *American Economic Review*, 104(5):431–35, May 2014b. doi: 10.1257/aer.104.5.431. URL https://www.aeaweb.org/articles?id=10.1257/aer.104.5.431.

Laurent, B. Adaptive estimation of a quadratic functional of a density by model selection. *Esaim: Probability and Statistics*, 9:1–18, 2005.

Luyo, P., Garcelon, E., Lazaric, A., and Pirotta, M. Differentially private exploration in reinforcement learning with linear representation. *arXiv preprint arXiv:2112.01585*, 2021.

Melo, F. S. and Ribeiro, M. I. $Q$ -learning with linear function approximation. In Bshouty, N. H. and Gentile, C. (eds.), *Learning Theory, 20th Annual Conference on Learning Theory, COLT 2007, San Diego, CA, USA, June 13-15, 2007, Proceedings*, volume 4539 of *Lecture Notes in Computer Science*, pp. 308–322. Springer, 2007. doi: 10.1007/978-3-540-72927-3_23. URL https://doi.org/10.1007/978-3-540-72927-3_23.

Mishra, N. and Thakurta, A. (nearly) optimal differentially private stochastic multi-arm bandits. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, UAI'15, pp. 592–601, Arlington, Virginia, USA, 2015. AUAI Press. ISBN 9780996643108.

Osband, I., Van Roy, B., and Wen, Z. Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*, pp. 2377–2386. PMLR, 2016.

Shariff, R. and Sheffet, O. Differentially private contextual linear bandits. *arXiv preprint arXiv:1810.00068*, 2018.

Tao, T. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

Tossou, A. C. Y. and Dimitrakakis, C. Achieving privacy in the adversarial multi-armed bandit. In *AAAI*, 2017.

Vietri, G., Balle, B., Krishnamurthy, A., and Wu, Z. S. Private reinforcement learning with pac and regret guarantees. *arXiv preprint arXiv:2009.09052*, 2020.

Wang, T., Zhou, D., and Gu, Q. Provably efficient reinforcement learning with linear function approximation under adaptivity constraints. *arXiv preprint arXiv:2101.02195*, 2021.

Zhang, F. *Matrix Theory: Basic Results and Techniques*. Springer, New York, NY, second edition, 2011. ISBN 9781461410997 1461410991 9781461410980 1461410983. doi: 10.1007/978-1-4614-1099-7.

Zhou, X. Differentially private reinforcement learning with linear function approximation. *CoRR*, abs/2201.07052, 2022. URL https://arxiv.org/abs/2201.07052.

# A. Privacy Analysis

**Proof of Theorem 14** Let $U$ and $U'$ be any two neighboring user sequences, and $\boldsymbol{y}_h^k, (\boldsymbol{y}')_h^k$ are the outputs of these two user sequences, respectively. Let $i \leq k$ be some episode with $x_h^i \neq x'^i_h$ and $a_h^i \neq a'^i_h$, where $(x_h^i, a_h^i) \in U$ and $(x'^i_h, a'^i_h) \in U'$. Then, by definition we have:

$$
\begin{aligned}
&\left\|\boldsymbol{y}_h^k - (\boldsymbol{y}')_h^k\right\|_2 \\
&= \left\| \sum_{j=1}^{k-1} \phi(x_h^j, a_h^j) \left( r_h(x_h^j, a_h^j) + V_{h+1}^j(x_{h+1}^j) \right) - \sum_{j=1}^{k-1} \phi(x'^j_h, a'^j_h) \left( r_h(x'^j_h, a'^j_h) + V_{h+1}^j(x'^j_{h+1}) \right) \right\|_2 \\
&= \left\| \phi(x_h^i, a_h^i) \left( r_h(x_h^i, a_h^i) + V_{h+1}^i(x_{h+1}^i) \right) - \phi(x'^i_h, a'^i_h) \left( r_h(x'^i_h, a'^i_h) + V_{h+1}^i(x'^i_{h+1}) \right) \right\|_2 \\
&\qquad \text{(since the two user sequences only differs in the } i\text{-th user.)} \\
&\leq 2 + \left\| \phi(x_h^i, a_h^i) V_{h+1}^i(x_{h+1}^i) - \phi(x'^i_h, a'^i_h) V_{h+1}^i(x'^i_{h+1}) \right\|_2 \\
&\qquad \text{(by linear MDP assumption, } r_h(x_h^i, a_h^i) \leq 1 \text{ and } \|\phi(x_h^i, a_h^i)\|_2 \leq 1) \\
&\leq 2 + \left\| \phi(x_h^i, a_h^i) V_{h+1}^i(x_{h+1}^i) \right\|_2 + \left\| \phi(x'^i_h, a'^i_h) V_{h+1}^i(x'^i_{h+1}) \right\|_2 \qquad \text{(by triangle inequality)} \\
&\leq 2 + \left| V_{h+1}^i(x_{h+1}^i) \right| + \left| V_{h+1}^i(x'^i_{h+1}) \right| \qquad \text{(by linear MDP assumption, } \|\phi(.,.)\|_2 \leq 1) \\
&\leq 2 + 2H
\end{aligned}
$$

Observe that for $\Delta(\boldsymbol{\Lambda}_h^k)$, by definition, we have

$$
\begin{aligned}
&\left\| \boldsymbol{\Lambda}_h^k - (\boldsymbol{\Lambda}')_h^k \right\| \\
&= \left\| \left( \lambda \boldsymbol{I} + \sum_{j=1}^{k-1} \phi(x_h^j, a_h^j) \phi(x_h^j, a_h^j)^\top \right) - \left( \lambda \boldsymbol{I} + \sum_{j=1}^{k-1} \phi(x'^j_h, a'^j_h) \phi(x'^j_h, a'^j_h)^\top \right) \right\|_2 \\
&= \left\| \phi(x_h^i, a_h^i) \phi(x_h^i, a_h^i)^\top - \phi(x'^i_h, a'^i_h) \phi(x'^i_h, a'^i_h)^\top \right\|_2 \qquad \text{(since the two user sequences only differ in the } i\text{-th user)} \\
&\leq \left\| \phi(x_h^i, a_h^i) \phi(x_h^i, a_h^i)^\top \right\|_2 + \left\| \phi(x'^i_h, a'^i_h) \phi(x'^i_h, a'^i_h)^\top \right\|_2 \qquad \text{(by triangle inequality)} \\
&\leq 2 \qquad \text{(Note that } \|\phi(x_h^t, a_h^t)\|_2 \leq 1)
\end{aligned}
$$

With these two bounds, we conclude our proof.

# B. Full Regret Analysis

## B.1. Notation and math tools

We start this section by listing our commonly used notations:

- $\lambda := 1$
- $\widetilde{\lambda}_y := \sqrt{\frac{\left(H \cdot \frac{dH}{\log 2} \log\left(1 + \frac{K}{\widetilde{\lambda}_\Lambda d}\right)\right)^2 \log\left(\frac{3KH}{p}\right)}{\rho}}$
- $\widetilde{\lambda}_\Lambda := \frac{\log(K)\left(6\sqrt{d+1} + 2\log\left(\frac{3KH}{p}\right)\right)}{\sqrt{2\rho}}$
- $U_K := \max\left\{1, 2H\sqrt{\frac{dK}{\widetilde{\lambda}_\Lambda}} + \frac{\widetilde{\lambda}_y}{\widetilde{\lambda}_\Lambda}\right\}$
- $\beta := 5H^2\sqrt{d\widetilde{\lambda}_\Lambda \log(\chi)} + 6dH\sqrt{\log(\chi)}$
- $\boldsymbol{M}_h^t := \widetilde{\boldsymbol{\Lambda}}_h^t - \boldsymbol{\Lambda}_h^t, \quad \boldsymbol{\eta}_h^k := \widetilde{\boldsymbol{y}}_h^k - \boldsymbol{y}_h^k$
- For any episode $k \in [K]$, $\tilde{k} < k$ is the last update episode before $k$.
- $\chi := \frac{25^4 \cdot 162 \cdot K^4 \cdot d \cdot U_K \cdot H}{p}$.

**Notation (as in (Jin et al., 2020))**

**Fact 17.** *For any two PSD matrices $A, B \in \mathbb{R}^{d\times d}$. If $B \succ A$ then for any vector $x \in \mathbb{R}^d$ we have $\|x\|_A \leq \|x\|_B$ and $\|x\|_{B^{-1}} \leq \|x\|_{B^{-1}}$.*

*Proof.* Since $A, B$ are PSD and $B \succ A$ then we have that $B - A \succ 0$ is also PSD. Then by definition of PSD we have $x^\top(B-A)x \geq 0$, which implies that $x^\top(B)x \geq x^\top(A)x$. □

### B.2. Private statistics

Recall from B.1 that $\boldsymbol{M}_h^k := \widetilde{\boldsymbol{\Lambda}}_h^k - \boldsymbol{\Lambda}_h^k$ and $\boldsymbol{\eta}_h^k := \widetilde{\boldsymbol{y}}_h^k - \boldsymbol{y}_h^k$. Then consider the following events:

$$\mathcal{E}_1 := \left\{\forall(k,h) \in [K]\times[H] : \left\|\boldsymbol{\eta}_h^k\right\|_2 \leq \widetilde{\lambda}_y\right\}, \quad \widetilde{\lambda}_y = \sqrt{\frac{\left(H\cdot\frac{dH}{\log 2}\log\left(1+\frac{K}{\widetilde{\lambda}_\Lambda d}\right)\right)^2 \log\left(\frac{3KH}{p}\right)}{\rho}} \tag{16}$$

$$\mathcal{E}_2 := \left\{\forall(k,h) \in [K]\times[H] : \left\|\boldsymbol{M}_h^k\right\|_2 \leq \widetilde{\lambda}_\Lambda\right\}, \quad \widetilde{\lambda}_\Lambda = \frac{\log(K)\left(6\sqrt{d+1}+2\log\left(\frac{3KH}{p}\right)\right)}{\sqrt{2\rho}} \tag{17}$$

$$\mathcal{E}_3 := \left\{\forall(k,h) \in [K]\times[H] : \left\|\sum_{i=1}^{\tilde{k}-1}\phi_h^i\left[V_{h+1}^{\tilde{k}}(x_{h+1}^i) - \mathbb{P}_h V_{h+1}^{\tilde{k}}(x_h^i, a_h^i)\right]\right\|_{(\widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}})^{-1}} \leq 6dH\sqrt{\log(\chi)}\right\} \tag{18}$$

Consider event $\mathcal{E} = \mathcal{E}_1 \cup \mathcal{E}_2 \cup \mathcal{E}_3$.

### B.3. High probability events

We will show that the probability of $\mathcal{E}$ is at most $p$. We begin by analyzing the error of the private algorithm. The next theorem states the utility guarantees given by the privacy mechanism:

**Theorem 18** (Utility)**.** *Let $\{\widetilde{\boldsymbol{\Lambda}}_h^k\}_{h\in[H],k\in[K]}$ and $\{\widetilde{\boldsymbol{y}}_h^k\}_{h\in[H],k\in[K]}$ be the privatized sufficient statistics of Algorithm 1 satisfying $\rho$-zCDP. Then with probability at least $1-p$ we have for any $k \in [K]$ and $h \in [H]$*

$$\|\widetilde{\boldsymbol{\Lambda}}_h^k - \boldsymbol{\Lambda}_h^k\|_{op} \leq \frac{16\sqrt{H}\log(K)\left(4\sqrt{d+1}+2\log\left(\frac{KH}{p}\right)\right)}{\sqrt{2\rho}}$$

*and for any $k \in [K]$ and $h \in [H]$ we have*

$$\|\widetilde{\boldsymbol{y}}_h^k - \boldsymbol{y}_h^k\|_2 \leq \sqrt{\frac{\left(\frac{dH}{\log 2}\log\left(1+\frac{K}{\widetilde{\lambda}_\Lambda d}\right)\right)^2\log\left(\frac{KH}{p}\right)}{\rho}}$$

By Theorem 18 we have that $\Pr[\mathcal{E}_1] \leq \frac{p}{3}$ and $\Pr[\mathcal{E}_2] \leq \frac{p}{3}$. From Theorem 25, condition on event $\mathcal{E}_1 \cup \mathcal{E}_2$, we have $\Pr[\mathcal{E}_3] \leq \frac{p}{3}$. Then by union bound, $\Pr[\mathcal{E}] \leq p$.

### B.4. Lemmas from Jin et al. (2020)

Here we state key lemmas in the regret analysis. See Appendix B.6 for the proofs.

**Lemma 19** (Lemma B.1 in (Jin et al., 2020))**.** *Under the linearity assumption, for any fixed policy $\pi$, let $\{\boldsymbol{w}_h^\pi\}_{h\in[H]}$ be the corresponding weights such that $Q_h^\pi(x,a) = \langle\phi(x,a), \boldsymbol{w}_h^\pi\rangle$ for all tuples $(x,a,h) \in \mathcal{S}\times\mathcal{A}\times[H]$. Then, we have*

$$\forall h \in [H], \|\boldsymbol{w}_h^\pi\| \leq 2H\sqrt{d}$$

**Lemma 20** (Lemma B.2 in (Jin et al., 2020))**.** *Under the event $\mathcal{E}$, for any $(k,h) \in [K]\times[H]$, the parameter vector $\boldsymbol{w}_h^k$ in Algorithm 1 satisfies:*

$$\left\|\boldsymbol{w}_h^k\right\| \leq U_K := \max\left\{1, 2H\sqrt{\frac{dK}{\widetilde{\lambda}_\Lambda}} + \frac{\widetilde{\lambda}_y}{\widetilde{\lambda}_\Lambda}\right\}$$

**Theorem 21** (Theorem D.3 in (Jin et al., 2020)). *Let $\{\epsilon_i\}_{i=1}^{\infty}$ be a real-valued stochastic process with corresponding filtration $\{\mathcal{F}_i\}_{i=0}^{\infty}$. Let $\epsilon_i|\mathcal{F}_{i=1}$ be zero-mean and $\sigma$-subGaussian; i.e. $\mathbb{E}[\epsilon_i|\mathcal{F}_{i-1}] = 0$, and*

$$\forall \lambda \in \mathbb{R}, \quad \mathbb{E}[\exp(\lambda \epsilon_i)|\mathcal{F}_{i-1}] \leq \exp(\lambda^2\sigma^2/2)$$

*Let $\{\phi_i\}_{i=0}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process where $\phi_i \in \mathcal{F}_{t-1}$. Assume $\Lambda_0$ is a $d \times d$ positive definite matrix, and let $\widetilde{\boldsymbol{\Lambda}}_i = \widetilde{\boldsymbol{\Lambda}}_0 + \sum_{s=1}^{i} \phi_s \phi_s^\top + \boldsymbol{M}_i$. Then for any $\delta > 0$, with probability at least $1-\delta$, we have for all $i \geq 0$:*

$$\left\| \sum_{s=1}^{i} \phi_s \epsilon_s \right\|^2_{(\widetilde{\boldsymbol{\Lambda}}^i)^{-1}} \leq 2\sigma^2 \log\left[ \frac{\det(\widetilde{\boldsymbol{\Lambda}}_i)^{1/2} \det(\widetilde{\boldsymbol{\Lambda}}_0)^{-1/2}}{\delta} \right]$$

**Lemma 22** (Lemma D.4 in (Jin et al., 2020)). *Let $\{x_i\}_{i=1}^{\infty}$ be a stochastic process on state space $\mathcal{S}$ with corresponding filtration $\{\mathcal{F}\}_{i=0}^{\infty}$. Let $\{\phi_i\}_{i=0}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process where $\phi_i \in \mathcal{F}_{i-1}$, and $\|\phi_i\| \leq 1$. Let $\widetilde{\boldsymbol{\Lambda}} = 2\widetilde{\lambda}_\Lambda \boldsymbol{I} + \boldsymbol{M} + \sum_{i=1}^{\tilde{k}} \phi_i \phi_i^\top$. Then for any $\delta > 0$, with probability at least $1-\delta$, for all $k \geq 0$, and any $V \in \mathcal{V}$ so that $\sup_x |V(x)| \leq H$, we have:*

$$\left\| \sum_{i=1}^{k} \phi_i \{V(x_i) - \mathbb{E}[V(x_i)|\mathcal{F}_{i-1}]\} \right\|^2_{(\widetilde{\boldsymbol{\Lambda}}^{\tilde{k}})^{-1}} \leq 4H^2 \left[ \frac{d}{2} \log\left( \frac{k + \widetilde{\lambda}_\Lambda}{\widetilde{\lambda}_\Lambda} + \log\left( \frac{\mathcal{N}_\epsilon}{\delta} \right) \right) \right] + \frac{8k^2\epsilon^2}{\widetilde{\lambda}_\Lambda}$$

*where $\mathcal{N}_\epsilon$ is the $\epsilon$-covering of $\mathcal{V}$ with respect to the distance $\operatorname{dist}(V, V') = \sup_x |V(x) - V'(x)|$.*

**Lemma 23** (Lemma D.5 in (Jin et al., 2020)). *For any $\epsilon > 0$, the $\epsilon$-covering number of the Euclidean ball in $\mathbb{R}^d$ with radius $R > 0$ is upper bounded by $(1 + 2R/\epsilon)^d$.*

**Lemma 24** (Similar to Lemma D.6 in (Jin et al., 2020)). *Let $\mathcal{V}$ denote a class of functions mappings from $\mathcal{S}$ to $\mathbb{R}$ with following parametric form:*

$$V(.) = \min\left\{ \max_a \boldsymbol{w}^\top \phi(., a) + \beta \sqrt{\phi(., a)^\top (\widetilde{\boldsymbol{\Lambda}})^{-1} \phi(., a)}, H \right\}$$

*where the parameters $(\boldsymbol{w}, \beta, \widetilde{\boldsymbol{\Lambda}})$ satisfy $\|\boldsymbol{w}\| \leq L, \beta \in [0, B]$ and the minimum eigenvalue satisfies $\lambda_{\min}(\widetilde{\boldsymbol{\Lambda}}) \geq \widetilde{\lambda}_\Lambda$. Assume $\|\phi(x, a)\| \leq 1$ for all $(x, a)$ pairs, and let $\mathcal{N}_\epsilon$ be the $\epsilon$-covering number of $\mathcal{V}$ with respect to the distance $\operatorname{dist}(V, V') = \sup_x |V(x) - V'(x)|$. Then,*

$$\log(\mathcal{N}_\epsilon) \leq d \log(1 + 4L/\epsilon) + d^2 \log(1 + 8d^{1/2}B^2/(\widetilde{\lambda}_\Lambda \epsilon)^2)$$

**Lemma 25** (Similar to Lemma B.3 in (Jin et al., 2020)). *Under the event $\mathcal{E}_1 \cup \mathcal{E}_2$, for $\chi = \frac{25^4 \cdot 162 \cdot K^4 \cdot d \cdot U_K \cdot H}{p}$ and $\beta = 5H^2\sqrt{d\widetilde{\lambda}_\Lambda \log(\chi)} + 6dH\sqrt{\log(\chi)}$, for any fixed $p \in (0, 1)$ if we let $\mathcal{E}_3$ be the event that*

$$\forall (k, h) \in [K] \times [H] : \left\| \sum_{i=1}^{\tilde{k}-1} \phi_h^i \left[ V_{h+1}^{\tilde{k}}(x_{h+1}^i) - \mathbb{P}_h V_{h+1}^{\tilde{k}}(x_h^i, a_h^i) \right] \right\|_{(\widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}})^{-1}} \leq 6dH\sqrt{\log(\chi)}$$

*Then $\mathbb{P}(\mathfrak{E}|\mathcal{E}) \geq 1 - \frac{p}{3}$*

Recall that our value function for episode $k$ and timestep $h$ is given by $Q_h^t(x, a) = \left\langle \phi(x, a), \boldsymbol{w}_h^{\tilde{k}} \right\rangle$, where $\tilde{k} < k$ is the last update episode.

**Lemma 26** (Similar to Lemma B.4 in (Jin et al., 2020)). *Under event $\mathcal{E}_1 \cup \mathcal{E}_2$, for any fixed policy $\pi$, on the event $\mathcal{E}_3$ defined in Theorem 25, we have for all $(x, a, h, k)$ that:*

$$Q_h^{\tilde{k}}(x, a) - Q_h^\pi(x, a) = \mathbb{P}_h \left( \boldsymbol{V}_{h+1}^{\tilde{k}} - \boldsymbol{V}_{h+1}^\pi \right)(x, a) + \Delta_h^k(x, a)$$

*for some $\Delta_h^k(x, a)$ that satisfies $|\Delta_h^k(x, a)| \leq \beta \|\phi(x, a)\|_{(\widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}})^{-1}}$.*

**Lemma 27** (Lemma B.5 in (Jin et al., 2020), (UCB)). *Under the event $\mathcal{E}$, we have that $Q_h^{\tilde{k}}(x,a) \geq Q_h^*(x,a)$ for all $(x,a,h,k) \in \mathcal{S} \times \mathcal{A} \times [H] \times [K]$.*

**Lemma 28** (Lemma B.6 in (Jin et al., 2020)(Recursive Lemma)). *Let $\tilde{k}$ be the last update episode before $k$. For any $(h,k) \in [H] \times [K]$, let $\tilde{\delta}_h^k = V_h^{\tilde{k}}(x_h^k) - V_h^{\pi_{\tilde{k}}}(x_h^k)$ denotes the errors of the estimated $V_h^{\tilde{k}}$ relative to $V_h^{\pi_{\tilde{k}}}$. Let $\zeta_{h+1}^k = \mathbb{E}[\tilde{\delta}_{h+1}^k | x_h^k, a_h^k] - \tilde{\delta}_{h+1}^k$. Then condition on the event $\mathfrak{E}$ defined in Theorem 25, we have the following: for any $(k,h) \in [K] \times [H]$*

$$\tilde{\delta}_h^k \leq \tilde{\delta}_{h+1}^k + \tilde{\zeta}_{h+1}^k + 2\beta \left\| (\phi_h^{\tilde{k}}) \right\|_{(\widetilde{\mathbf{\Lambda}}_h^{\tilde{k}})^{-1}}$$

### B.5. Final Regret Proof

Finally, we prove the regret bound stated in Theorem 16. For any $k \in [K]$, let $\tilde{k} \leq k$ be the last episode the algorithm updated its policy. That is, on episode $k$ the agent acts according to policy $\pi_k$ which is given by $\pi_k(x,h) = \arg\max_a Q_h^{\tilde{k}}(x,a)$.

We begin by decomposing the regret using the optimistic approximation. Thus, by Theorem 27 we can decompose the regret as follows:

$$R(K) = \sum_{k=1}^{K} [V_1^*(x_1^k) - V_1^{\pi_k}(x_1^k)] \leq \sum_{k=1}^{K} [V_1^{\tilde{k}}(x_1^k) - V_1^{\pi_k}(x_1^k)] = \sum_{k=1}^{K} \tilde{\delta}_1^k$$

Recall that $\tilde{\delta}_h^k = V_h^{\tilde{k}}(x_h^k) - V_h^{\pi_k}(x_h^k)$ and $\zeta_{h+1}^k = \mathbb{E}[\tilde{\delta}_{h+1}^k | x_h^k, a_h^k] - \tilde{\delta}_{h+1}^k$. On the event $\mathfrak{E}$, by the recursive lemma Theorem 28 we obtain the following:

$$R(K) \leq \sum_{k=1}^{K} \sum_{h=1}^{H} \tilde{\zeta}_h^k + 3\beta \sum_{k=1}^{K} \sum_{h=1}^{H} \sqrt{(\phi_h^k)^\top \left( \widetilde{\mathbf{\Lambda}}_h^{\tilde{k}} \right)^{-1} \phi_h^k}$$

For the first term, since the computation of $V_h^{\tilde{k}}$ is independent of the new observation $x_h^k$ at episode $k$, we obtain that $\{\zeta_h^k\}$ is a martingale difference sequence satisfying $\zeta_h^k \leq 2H$ for all $k,h \in [K] \times [H]$. Therefore, we can apply the Azuma-Hoeffding inequality, for any $t > 0$, we have :

$$\Pr\left[ \sum_{k=1}^{K} \sum_{h=1}^{H} \tilde{\zeta}_h^k > t \right] \leq \exp\left( \frac{-t^2}{2KH^3} \right)$$

Hence, with probability at least $1 - p/4$, we have

$$\sum_{k=1}^{K} \sum_{h=1}^{H} \tilde{\zeta}_h^k \leq \sqrt{2KH^3 \log(4/p)} \leq \sqrt{2KH^3 \log(\chi)} \tag{19}$$

where $\chi := \frac{24^2 \cdot 18 \cdot K^2 d U_K H}{p}$. On the event $\mathcal{E}_2$, the condition of Theorem 11 is satisfied, which means that the total number of updates is bounded by $\frac{dH}{\log 2} \log\left(1 + \frac{K}{\widetilde{\lambda}_\Lambda d}\right)$ and the condition $N_{\text{count}} > N_{\max}$ from Line 8 in Algorithm 1 never happens. It follows that on $\mathcal{E}_2$, the following bound holds: .

$$(\phi_h^k)^\top \left( \widetilde{\mathbf{\Lambda}}_h^{\tilde{k}} \right)^{-1} \phi_h^k \leq 2(\phi_h^k)^\top \left( \widetilde{\mathbf{\Lambda}}_h^{k} \right)^{-1} \phi_h^k \text{ for all } (k,h) \in [K] \times [H] \tag{20}$$

Thus, by Equation (19) and Equation (20) we have

$$R(K) \leq \sqrt{2KH^3 \log(\chi)} + 6\beta \sum_{k=1}^{K} \sum_{h=1}^{H} \sqrt{(\phi_h^k)^\top \left( \widetilde{\mathbf{\Lambda}}_h^{k} \right)^{-1} \phi_h^k}$$

The next step is to apply Cauchy-Schwarz inequality to the second term:

$$R(K) \leq \sqrt{2KH^3 \log(\chi)} + 6\beta \sum_{h=1}^{H} \sqrt{K} \sqrt{\sum_{k=1}^{K} (\phi_h^k)^\top (\widetilde{\mathbf{\Lambda}}_h^k)^{-1} \phi_h^k}$$

Now recall that $\widetilde{\mathbf{\Lambda}}_h^k = \mathbf{\Lambda}_h^k + 2\widetilde{\lambda}_\Lambda \boldsymbol{I} + \boldsymbol{M}_h^k$. We use the fact that on event $\mathcal{E}_2$ (Equation (17)), we have $\|\boldsymbol{M}_h^k\| \leq \widetilde{\lambda}_\Lambda$ and thus $\boldsymbol{M}_h^k + \widetilde{\lambda}_\Lambda \boldsymbol{I} \succ 0$, which implies that the minimum eigenvalue of $\widetilde{\mathbf{\Lambda}}_h^k$ is $\widetilde{\lambda}_\Lambda$. Then by Theorem 21 from (Jin et al., 2020) and $\|\mathbf{\Lambda}_h^K\| \leq \widetilde{\lambda}_\Lambda + K$ we have that

$$\sum_{k=1}^{K} (\phi_h^k)^\top \left(\widetilde{\mathbf{\Lambda}}_h^k\right)^{-1} \phi_h^k \leq 2 \log \left( \frac{\det \widetilde{\mathbf{\Lambda}}_h^{K+1}}{\det \widetilde{\mathbf{\Lambda}}_h^1} \right) \leq 2d \log \left( \frac{\widetilde{\lambda}_\Lambda + K}{\widetilde{\lambda}_\Lambda} \right) \leq 2d \log(\chi) \tag{21}$$

Finally, by Equation (21), and by substituting $\beta = 5H^2 \sqrt{d\widetilde{\lambda}_\Lambda \log(\chi)} + 6dH\sqrt{\log(\chi)}$, we can bound the total regret as:

$$\begin{aligned}
&R(K) \\
&\leq \sqrt{2KH^3 \log(\chi)} + 6\beta H \sqrt{2dK \log(\chi)} \\
&\leq \sqrt{2KH^3 \log(\chi)} + 36H^2 d^{3/2} \sqrt{2K \log(\chi)^2} + 30H^3 d \sqrt{2K\widetilde{\lambda}_\Lambda \log(\chi)^2}
\end{aligned}$$

Recall that $\widetilde{\lambda}_\Lambda = \frac{\log(K)\left(6\sqrt{d+1} + 2\log\left(\frac{3KH}{p}\right)\right)}{\sqrt{2\rho}}$. Therefore, by a union bound on the events $\mathfrak{E}, \mathcal{E}_1, \mathcal{E}_2$, and the event from Equation (19) we have that with probability at least $1-p$, the regret of Algorithm 1 is bounded by

$$R(K) = \widetilde{O}\left(d^{3/2}H^2K^{1/2} + H^3 d^{5/4} K^{1/2} \rho^{-1/4}\right)$$

### B.6. Lemma proofs from Jin et al. (2020)

**Proof of Theorem 20 (Lemma B.2 in Jin et al. (2020))** For any unit vector $v \in \mathbb{R}^d$, we have:

$$\begin{aligned}
|v^\top \boldsymbol{w}_h^k| &= \left| v^\top \left(\widetilde{\mathbf{\Lambda}}_h^{\tilde{k}}\right)^{-1} \widetilde{\boldsymbol{y}}_k^{\tilde{k}} \right| \\
&= \left| v^\top \left(\widetilde{\mathbf{\Lambda}}_h^{\tilde{k}}\right)^{-1} \left( \sum_{i=1}^{\tilde{k}-1} \phi_h^i [r(x_h^i, a_h^i) + \max_a Q_{h+1}(x_{h+1}^i, a)] + \boldsymbol{\eta}_h^{\tilde{k}} \right) \right| \\
&\leq \sum_{i=1}^{\tilde{k}-1} \left| v^\top \left(\widetilde{\mathbf{\Lambda}}_h^{\tilde{k}}\right)^{-1} \phi_h^i \right| (H+1) + \left| v^\top \left(\widetilde{\mathbf{\Lambda}}_h^{\tilde{k}}\right)^{-1} \boldsymbol{\eta}_h^{\tilde{k}} \right| \\
&\leq \sqrt{\left[ \sum_{i=1}^{\tilde{k}-1} v^\top \left(\widetilde{\mathbf{\Lambda}}_h^{\tilde{k}}\right)^{-1} v \right] \left[ \sum_{i=1}^{\tilde{k}-1} (\phi_h^i)^\top \left(\widetilde{\mathbf{\Lambda}}_h^{\tilde{k}}\right)^{-1} \phi_h^i \right]} 2H + \|v\| \left\| \left(\widetilde{\mathbf{\Lambda}}_h^k\right)^{-1} \right\| \left\| \boldsymbol{\eta}_h^{\tilde{k}} \right\| \\
&\leq 2H \sqrt{\frac{d(\tilde{k}-1)}{\widetilde{\lambda}_\Lambda}} + \frac{\widetilde{\lambda}_y}{\widetilde{\lambda}_\Lambda} \\
&\leq 2H \sqrt{\frac{dK}{\widetilde{\lambda}_\Lambda}} + \frac{\widetilde{\lambda}_y}{\widetilde{\lambda}_\Lambda} \leq U_K
\end{aligned}$$

**Proof of Theorem 22 (Lemma D.4 in Jin et al. (2020))** For any $V \in \mathcal{V}$, we know there exists a $\tilde{V}$ in the $\epsilon$-covering such that

$$V = \tilde{V} + \Delta_V \quad \text{and} \quad \sup_x |\Delta_V(x)| \leq \epsilon$$

This gives the following decomposition:

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

where we can apply Theorem 21 and a union bound to the first term, and bound the second term by $8k^2\epsilon^2/\widetilde{\lambda}_\Lambda$.

**Proof of Theorem 24 (Lemma D.6 in Jin et al. (2020))** Equivalently, we can reparametrize the function class $\mathcal{V}$ by letting $A = \beta^2(\widetilde{\mathbf{\Lambda}})^{-1}$, so we have

$$V(.) = \min\{\max_a \boldsymbol{w}^\top \phi(., a) + \sqrt{\phi(., a)^\top A \phi(., a)}, H\} \tag{22}$$

for $\|\boldsymbol{w}\| \leq L$ and $\|A\| \leq B^2(\lambda + \widetilde{\mathbf{\Lambda}}_\Lambda)^{-1}$. For any two functions $V_1, V_2 \in \mathcal{V}$, let them take the form in Equation (22) with parameters $(\boldsymbol{w}_1, A_1)$ and $(\boldsymbol{w}_2, A_2)$, respectively. Then, since both $\min\{., H\}$ and $\max_a$ are contraction maps, we have:

$$\begin{aligned}
\text{dist}(V_1, V_2) &\leq \sup_{x,a} \left| \left[ \boldsymbol{w}_1^\top \phi(x, a) + \sqrt{\phi(x, a)^\top A_2 \phi(x, a)} \right] - \left[ \boldsymbol{w}_2^\top \phi(x, a) + \sqrt{\phi(x, a)^\top A_2 \phi(x, a)} \right] \right| \\
&\leq \sup_{\phi: \|\phi\| \leq 1} \left| \left[ \boldsymbol{w}_1^\top \phi + \sqrt{\phi^\top A_2 \phi} \right] - \left[ \boldsymbol{w}_2^\top \phi + \sqrt{\phi^\top A_2 \phi} \right] \right| \\
&\leq \sup_{\phi: \|\phi\| \leq 1} \left| (\boldsymbol{w}_1 - \boldsymbol{w}_2)^\top \phi \right| + \sup_{\phi: \|\phi\| \leq 1} \sqrt{|\phi^\top (A_1 - A_2) \phi|} \\
&= \|\boldsymbol{w}_1 - \boldsymbol{w}_2\| + \sqrt{\|A_1 - A_2\|} \\
&\leq \|\boldsymbol{w}_1 - \boldsymbol{w}_2\| + \sqrt{\|A_1 - A_2\|_F}
\end{aligned} \tag{23}$$

where the second last inequality follows from the fact that $|\sqrt{x} - \sqrt{y}| \leq \sqrt{|x - y|}$ holds for any $x, y \geq 0$. For matrices, $\|.\|$ and $\|.\|_F$ denote the matrix operator norm and Frobenius norm respectively.

Let $\mathcal{C}_{\boldsymbol{w}}$ be an $\epsilon/2$-cover of $\{\boldsymbol{w} \in \mathbb{R}^d | \, \|\boldsymbol{w}\| \leq L\}$ with respect to the 2-norm, and $\mathcal{C}_A$ be an $\epsilon^2/4$-cover of $\{A \in \mathbb{R}^{d \times d} | \, \|A\|_F \leq d^{1/2} B^2 \widetilde{\lambda}_\Lambda^{-1}\}$ with respect to the Frobenius norm. By Theorem 23, we know:

$$|\mathcal{C}_{\boldsymbol{w}}| \leq (1 + 4L/\epsilon)^d, \quad |\mathcal{C}_A| \leq [1 + 8d^{1/2} B^2/(\widetilde{\lambda}_\Lambda \epsilon^2)]^{d^2}$$

By Equation (23), for any $V_1 \in \mathcal{V}$, there exists $\boldsymbol{w}_2 \in \mathcal{C}_{\boldsymbol{w}}$ and $A_2 \in \mathcal{C}_A$ such that $V_2$ parametrized by $(\boldsymbol{w}_2, A_2)$ satisfies $\text{dist}(V_1, V_2) \leq \epsilon$. Hence, it holds that $\mathcal{N}_\epsilon \leq |\mathcal{C}_{\boldsymbol{w}}||\mathcal{C}_A|$, which gives:

$$\log \mathcal{N}_\epsilon \leq \log |\mathcal{C}_{\boldsymbol{w}}| + \log |\mathcal{C}_A| \leq d \log(1 + 4L/\epsilon) + d^2 \log[1 + 8d^{1/2} B^2/(\widetilde{\lambda}_\Lambda \epsilon^2)]$$

This concludes the proof.

**Proof of Theorem 25 (Lemma B.3 in Jin et al. (2020))** For all $(k, h) \in [K] \times [H]$, by Theorem 20, we have $\left\|w_h^k\right\| \leq U_K$. In addition, by construction of $\widetilde{\mathbf{\Lambda}}_h^{\tilde{k}}$, the minimum eigenvalue of $\widetilde{\mathbf{\Lambda}}_h^{\tilde{k}}$ is lowered bounded by $\widetilde{\lambda}_\Lambda$. Thus, by combining

Theorem 22 and Theorem 24, with probability at least $1 - \frac{p}{6H}$,we have for any $k > 1$ that:

$$\left\| \sum_{i=1}^{\tilde{k}-1} \phi_h^i [V_{h+1}^{\tilde{k}}(x_{h+1}^i) - \mathbb{P}_h V_{h+1}^{\tilde{k}}(x_h^i, a_h^i)] \right\|^2_{(\widetilde{\mathbf{\Lambda}}_h^{\tilde{k}})^{-1}} \tag{24}$$

$$\leq 4H^2 \left[ \frac{d}{2} \log\left( \frac{k + \widetilde{\lambda}_\Lambda}{\widetilde{\lambda}_\Lambda} \right) + d \log\left( 1 + \frac{4U_K}{\epsilon_0} \right) + d^2 \log\left( 1 + \frac{8d^{1/2}\beta^2}{\epsilon_0^2 \widetilde{\lambda}_\Lambda} \right) + \log\left( \frac{6H}{p} \right) \right] + \frac{8k^2\epsilon_0^2}{\widetilde{\lambda}_\Lambda} \tag{25}$$

with $\epsilon_0 = \frac{dH}{k}\sqrt{\widetilde{\lambda}_\Lambda}$. Notice that we choose the hyperparameter $\beta = 5H^2\sqrt{d\widetilde{\lambda}_\Lambda \log(\chi)} + 6dH\sqrt{\log(\chi)}$. We have:

$$\left\| \sum_{i=1}^{\tilde{k}-1} \phi_h^i [V_{h+1}^{\tilde{k}}(x_{h+1}^i) - \mathbb{P}_h V_{h+1}^{k}(x_h^i, a_h^i)] \right\|^2_{(\widetilde{\mathbf{\Lambda}}_h^{k})^{-1}}$$

$$\leq \underbrace{4H^2d^2 \left( 2 + \log\left( 1 + \frac{8.25^2K^2\sqrt{d}\log(\chi)}{\widetilde{\lambda}_\Lambda} \right) \right)}_{G_1} + \underbrace{2H^2 d \log\left( 1 + \frac{K}{\widetilde{\lambda}_\Lambda} \right)}_{G_2}$$

$$+ \underbrace{4H^2 d \log\left( 1 + \frac{4U_K K}{dH\sqrt{\widetilde{\lambda}_\Lambda}} \right)}_{G_3} + \underbrace{4H^2 \log\left( \frac{6H}{p} \right)}_{G_4}$$

We can bound each term individually:

- $G_1 \leq 4H^2d^2[2 + \log(1 + 8 \cdot 25^2K^2\sqrt{d}\log(\chi))]$
  First, we have $\log(1 + 8 \cdot 24^2K^2\log^2(\chi)) \leq 9 \cdot 25^2K^2\sqrt{d}\log(\chi)$.

  Notice that if we set $\chi \geq (9 \cdot 25^2K^2\sqrt{d})(18 \cdot 25^2K^2\sqrt{d}) = 9 \cdot 18 \cdot 25^4K^4d$, then we can upper bound $\chi \geq 9 \cdot 25^2K^2\sqrt{d}\log(\chi)$. Hence, we can write:

  $$G_1 \leq 8H^2d^2\log(\chi) + 4H^2d^2\log(\chi) = 16H^2d^2\log(\chi)$$

- We have $G_2 \leq 2H^2d^2\log(2K) \leq 2H^2d^2\log(\chi)$ since $\chi \geq 2K$.
- We have $G_3 = 4H^2d\log\left(1 + \frac{4U_KK}{dH\sqrt{\widetilde{\lambda}_\Lambda}}\right) \leq 4H^2d^2\log(5U_KK) \leq 4H^2d^2\log(\chi)$ since $\chi \geq 5U_KK$
- We have $G_4 \leq 4H^2d^2\log(\chi)$

Therefore,

$$\left\| \sum_{i=1}^{\tilde{k}-1} \phi_h^i \left[ V_{h+1}^{k}(x_{h+1}^i) - \mathbb{P}_h V_{h+1}^{k}(x_h^i, a_h^i) \right] \right\|^2_{(\widetilde{\mathbf{\Lambda}}_h^{\tilde{k}})^{-1}} \leq 26d^2H^2\log(\chi)$$

**Proof of Theorem 26 (Lemma B.4 in Jin et al. (2020))** By Theorem 2 and the Bellman equation, we have that for any tuple $(x, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$:

$$Q_h^\pi(x, a) := \langle \phi(x, a), \boldsymbol{w}_h^\pi \rangle = (r_h + \mathbb{P}_h V_{h+1}^\pi)(x, a)$$

For any $k \in [K]$, the action-value function in Algorithm 1 is defined as

$$Q_h^k(x, a) = \left\langle \phi(x, a), \boldsymbol{w}_h^{\tilde{k}} \right\rangle$$

where $\boldsymbol{w}_h^{\tilde{k}} = \left(\widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}}\right)^{-1} \widetilde{\boldsymbol{y}}_h^{\tilde{k}}$ and

$$\widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}} = 2\widetilde{\lambda}_\Lambda \boldsymbol{I} + \boldsymbol{M}_h^{\tilde{k}} + \sum_{i=1}^{\tilde{k}-1} \phi\left(x_h^i, a_h^i\right) \phi\left(x_h^i, a_h^i\right)^\top$$
$$\widetilde{\boldsymbol{y}}_h^{\tilde{k}} = \sum_{i=1}^{\tilde{k}-1} \phi\left(x_h^i, a_h^i\right) \left(r_h(x_h^i, a_h^i) + V_{h+1}^{\tilde{k}}(x_{h+1}^i)\right) + \boldsymbol{\eta}_h^{\tilde{k}}$$

Note that $\tilde{k} \leq k$ is the last update episode for any $h \in [H]$ before episode $k$ and under event $\mathcal{E}$, the minimum eigenvalue of $\widetilde{\boldsymbol{\Lambda}}_h^k$ is $\widetilde{\lambda}_\Lambda$. This means that

$$\|v\|_{(\widetilde{\boldsymbol{\Lambda}}_h^k)^{-1}} \leq \|v\|_{((\widetilde{\lambda}_\Lambda)\boldsymbol{I})^{-1}} = \sqrt{v^\top (\widetilde{\lambda}_\Lambda \boldsymbol{I})^{-1} v} \leq \sqrt{\frac{1}{\widetilde{\lambda}_\Lambda}} \|v\| \tag{26}$$

Hence, we begin by decomposing the term $\boldsymbol{w}_h^{\tilde{k}} - \boldsymbol{w}_h^\pi$ as

$$\begin{aligned}
&\boldsymbol{w}_h^{\tilde{k}} - \boldsymbol{w}_h^\pi \\
&= (\widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}})^{-1} \left( \sum_{i=1}^{\tilde{k}-1} \phi_h^i [r_h^i + V_{h+1}^{\tilde{k}}(x_{h+1}^i)] + \boldsymbol{\eta}_h^{\tilde{k}} \right) - \boldsymbol{w}_h^\pi \\
&= (\widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}})^{-1} \left( \sum_{i=1}^{\tilde{k}-1} \phi_h^i [r_h^i + V_{h+1}^{\tilde{k}}(x_{h+1}^i)] \right) - \boldsymbol{w}_h^\pi + \left(\widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}}\right)^{-1} \boldsymbol{\eta}_h^{\tilde{k}} \\
&= (\widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}})^{-1} \left( \sum_{i=1}^{\tilde{k}-1} \phi_h^i r_h^i + \sum_{i=1}^{\tilde{k}-1} \phi_h^i V_{h+1}^{\tilde{k}}(x_{h+1}^i) + \phi_h^i \mathbb{P}_h V_{h+1}^\pi(x_h^i, a_h^i) - \phi_h^i \mathbb{P}_h V_{h+1}^\pi(x_h^i, a_h^i) \right) \\
&\quad - \boldsymbol{w}_h^\pi + \left(\widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}}\right)^{-1} \boldsymbol{\eta}_h^{\tilde{k}} \\
&= (\widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}})^{-1} \left( \sum_{i=1}^{\tilde{k}-1} \phi_h^i \left(r_h^i + \mathbb{P}_h V_{h+1}^\pi(x_h^i, a_h^i)\right) + \sum_{i=1}^{\tilde{k}-1} \phi_h^i \left(V_{h+1}^{\tilde{k}}(x_{h+1}^i) - \mathbb{P}_h V_{h+1}^\pi(x_h^i, a_h^i)\right) \right) \\
&\quad - \boldsymbol{w}_h^\pi + \left(\widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}}\right)^{-1} \boldsymbol{\eta}_h^{\tilde{k}}
\end{aligned}$$

By definition, we can replace $\left\langle \phi_h^i, \boldsymbol{w}_h^\pi \right\rangle = r_h^i + \mathbb{P}_h V_{h+1}^\pi(x_h^i, a_h^i)$. Then, we can continue expanding the equality above as:

$$\begin{aligned}
&\boldsymbol{w}_h^{\tilde{k}} - \boldsymbol{w}_h^\pi \\
&= (\widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}})^{-1} \left( \sum_{i=1}^{\tilde{k}-1} \phi_h^i \langle \phi_h^i, \boldsymbol{w}_h^\pi \rangle + \sum_{i=1}^{\tilde{k}-1} \phi_h^i \left(V_{h+1}^{\tilde{k}}(x_{h+1}^i) - \mathbb{P}_h V_{h+1}^\pi(x_h^i, a_h^i)\right) \right) - \boldsymbol{w}_h^\pi + \left(\widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}}\right)^{-1} \boldsymbol{\eta}_h^{\tilde{k}} \\
&= (\widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}})^{-1} \left( \sum_{i=1}^{\tilde{k}-1} \phi_h^i \langle \phi_h^i, \boldsymbol{w}_h^\pi \rangle - \widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}} \boldsymbol{w}_h^\pi + \sum_{i=1}^{\tilde{k}-1} \phi_h^i \left(V_{h+1}^{\tilde{k}}(x_{h+1}^i) - \mathbb{P}_h V_{h+1}^\pi(x_h^i, a_h^i)\right) \right) + \left(\widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}}\right)^{-1} \boldsymbol{\eta}_h^{\tilde{k}}
\end{aligned}$$

Then, plugging in for the definition of $\widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}}$ we get

$$
\begin{aligned}
&= (\widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}})^{-1} \left( \sum_{i=1}^{\tilde{k}-1} \phi_h^i \langle \phi_h^i, \boldsymbol{w}_h^\pi \rangle - \left( 2\widetilde{\lambda}_\Lambda \boldsymbol{I} + \boldsymbol{M}_h^{\tilde{k}} + \sum_{i=1}^{\tilde{k}-1} \phi_h^i {\phi_h^i}^\top \right) \boldsymbol{w}_h^\pi \right) \\
&\quad + (\widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}})^{-1} \left( \sum_{i=1}^{\tilde{k}-1} \phi_h^i \left( V_{h+1}^{\tilde{k}}(x_{h+1}^i) - \mathbb{P}_h V_{h+1}^\pi(x_h^i, a_h^i) \right) \right) + \left( \widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}} \right)^{-1} \boldsymbol{\eta}_h^{\tilde{k}} \\
&= (\widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}})^{-1} \left\{ \left( -2\widetilde{\lambda}_\Lambda \boldsymbol{I} - \boldsymbol{M}_h^{\tilde{k}} \right) \boldsymbol{w}_h^\pi + \sum_{i=1}^{\tilde{k}-1} \phi_h^i [V_{h+1}^{\tilde{k}}(x_{h+1}^i) - \mathbb{P}_h V_{h+1}^\pi(x_h^i, a_h^i)] \right\} + \left( \widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}} \right)^{-1} \boldsymbol{\eta}_h^{\tilde{k}} \\
&= \underbrace{-(\widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}})^{-1} \left( 2\widetilde{\lambda}_\Lambda \boldsymbol{I} + \boldsymbol{M}_h^{\tilde{k}} \right) \boldsymbol{w}_h^\pi}_{\mathbf{q}_1} + \underbrace{(\widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}})^{-1} \sum_{i=1}^{\tilde{k}-1} \phi_h^i [V_{h+1}^{\tilde{k}}(x_{h+1}^i) - \mathbb{P}_h V_{h+1}^{\tilde{k}}(x_h^i, a_h^i)]}_{\mathbf{q}_2} + \\
&\quad + \underbrace{(\widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}})^{-1} \sum_{i=1}^{\tilde{k}-1} \phi_h^i \mathbb{P}_h (V_{h+1}^{\tilde{k}} - V_{h+1}^\pi)(x_h^i, a_h^i)}_{\mathbf{q}_3} + \underbrace{\left( \widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}} \right)^{-1} \boldsymbol{\eta}_h^{\tilde{k}}}_{\mathbf{q}_4}
\end{aligned}
$$

We proceed to bound each term in the right hand side individually. For the first term, we have:

$$
\begin{aligned}
|\langle \phi(x,a), \mathbf{q}_1 \rangle| &= \left\langle \phi(x,a), (\widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}})^{-1} \left( 2\widetilde{\lambda}_\Lambda \boldsymbol{I} + \boldsymbol{M}_h^{\tilde{k}} \right) \boldsymbol{w}_h^\pi \right\rangle \\
&= \left\langle \phi(x,a) (\widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}})^{-1/2}, (\widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}})^{-1/2} \left( 2\widetilde{\lambda}_\Lambda \boldsymbol{I} + \boldsymbol{M}_h^{\tilde{k}} \right) \boldsymbol{w}_h^\pi \right\rangle \\
&\leq \|\phi(x,a)\|_{(\widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}})^{-1}} \left\| \left( 2\widetilde{\lambda}_\Lambda \boldsymbol{I} + \boldsymbol{M}_h^{\tilde{k}} \right) \boldsymbol{w}_h^\pi \right\|_{(\widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}})^{-1}} \\
&\leq \|\phi(x,a)\|_{(\widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}})^{-1}} \left\| \left( \widetilde{\lambda}_\Lambda \boldsymbol{I} \right) \boldsymbol{w}_h^\pi \right\|_{(\widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}})^{-1}} \\
&\leq \frac{\widetilde{\lambda}_\Lambda}{\sqrt{\widetilde{\lambda}_\Lambda}} \|w_h^\pi\| \, \|\phi(x,a)\|_{(\widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}})^{-1}} && \text{(by Equation (26))} \\
&= 2H\sqrt{d\widetilde{\lambda}_\Lambda} \, \|\phi(x,a)\|_{(\widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}})^{-1}} && \text{(by Theorem 19)}
\end{aligned}
$$

As a straightforward application of Theorem 25, we can bound the second term as such:

$$
\begin{aligned}
|\langle \phi(x,a), \mathbf{q}_2 \rangle| &\leq 6dH\sqrt{\log(\chi)} \sqrt{\phi(x,a)^\top (\widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}})^{-1} \phi(x,a)} \\
&= 6dH\sqrt{\log(\chi)} \, \|\phi(x,a)\|_{(\widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}})^{-1}}
\end{aligned}
$$

For the third term, by definition of linear MDP, we have:

$$\begin{aligned}
\langle \phi(x,a), \mathbf{q}_3 \rangle &= \left\langle \phi(x,a), (\widetilde{\mathbf{\Lambda}}_h^{\tilde{k}})^{-1} \sum_{i=1}^{\tilde{k}-1} \phi_h^i \mathbb{P}_h (V_{h+1}^{\tilde{k}} - V_{h+1}^{\pi})(x_h^i, a_h^i) \right\rangle \\
&= \left\langle \phi(x,a), (\widetilde{\mathbf{\Lambda}}_h^{\tilde{k}})^{-1} \sum_{i=1}^{\tilde{k}-1} \phi_h^i \phi_t^{i\top} \int (V_{h+1}^{\tilde{k}} - V_{h+1}^{\pi})(x') d\mu_h(x') \right\rangle \\
&= \underbrace{\left\langle \phi(x,a), \int (V_{h+1}^{\tilde{k}} - V_{h+1}^{\pi})(x') d\mu_h(x') \right\rangle}_{p_1} \\
&\quad \underbrace{- \left\langle \phi(x,a), (\widetilde{\mathbf{\Lambda}}_h^{\tilde{k}})^{-1} \left( 2\widetilde{\lambda}_\Lambda \boldsymbol{I} + \boldsymbol{M}_h^{\tilde{k}} \right) \int (V_{h+1}^{\tilde{k}} - V_{h+1}^{\pi})(x') d\mu_h(x') \right\rangle}_{p_2}
\end{aligned}$$

where, by assumption, we have

$$p_1 = \mathbb{P}_h (V_{h+1}^{\tilde{k}} - V_{h+1}^{\pi})(x,a)$$

and similar to $\mathbf{q}_1$, we can bound $|p_2|$ as follows:

$$\begin{aligned}
|p_2| &\le \sqrt{\phi(x,a)^\top (\widetilde{\mathbf{\Lambda}}_h^{\tilde{k}})^{-1} \phi(x,a)} \left\| (\widetilde{\mathbf{\Lambda}}_h^{\tilde{k}})^{-1/2} \right\| \left\| (2\widetilde{\lambda}_\Lambda) \boldsymbol{I} + \boldsymbol{M}_h^{\tilde{k}} \right\| 2H\sqrt{d} \\
&\le \frac{1}{\sqrt{\widetilde{\lambda}_\Lambda}} (\widetilde{\lambda}_\Lambda) 2H\sqrt{d} \sqrt{\phi(x,a)^\top (\widetilde{\mathbf{\Lambda}}_h^{\tilde{k}})^{-1} \phi(x,a)} \\
&\le 2H\sqrt{d\widetilde{\lambda}_\Lambda} \left\| \phi(x,a) \right\|_{(\widetilde{\mathbf{\Lambda}}_h^{\tilde{k}})^{-1}}
\end{aligned}$$

For the fourth term, we have:

$$\begin{aligned}
|\langle \phi(x,a), \mathbf{q}_4 \rangle| &\le \left| \left\langle \phi(x,a), \left( \widetilde{\mathbf{\Lambda}}_h^{\tilde{k}} \right)^{-1} \boldsymbol{\eta}_h^{\tilde{k}} \right\rangle \right| \\
&= \left| \phi(x,a)^\top (\widetilde{\mathbf{\Lambda}}_h^{\tilde{k}})^{-1/2} (\widetilde{\mathbf{\Lambda}}_h^{\tilde{k}})^{-1/2} \boldsymbol{\eta}_h^{\tilde{k}} \right| \\
&= \left| \left( (\widetilde{\mathbf{\Lambda}}_h^{\tilde{k}})^{-1/2} \phi(x,a) \right)^\top \left( (\widetilde{\mathbf{\Lambda}}_h^{\tilde{k}})^{-1/2} \boldsymbol{\eta}_h^{\tilde{k}} \right) \right| \\
&\le \left\| (\widetilde{\mathbf{\Lambda}}_h^{\tilde{k}})^{-1/2} \phi(x,a) \right\| \left\| (\widetilde{\mathbf{\Lambda}}_h^{\tilde{k}})^{-1/2} \boldsymbol{\eta}_h^{\tilde{k}} \right\| \\
&= \left\| (\widetilde{\mathbf{\Lambda}}_h^{\tilde{k}})^{-1/2} \boldsymbol{\eta}_h^{\tilde{k}} \right\| \sqrt{\phi(x,a)^\top (\widetilde{\mathbf{\Lambda}}_h^{\tilde{k}})^{-1} \phi(x,a)} \\
&= \left\| \boldsymbol{\eta}_h^{\tilde{k}} \right\|_{(\widetilde{\mathbf{\Lambda}}_h^{\tilde{k}})^{-1}} \sqrt{\phi(x,a)^\top (\widetilde{\mathbf{\Lambda}}_h^{\tilde{k}})^{-1} \phi(x,a)} \\
&\le \frac{\left\| \boldsymbol{\eta}_h^{\tilde{k}} \right\|}{\sqrt{\widetilde{\lambda}_\Lambda}} \sqrt{\phi(x,a)^\top (\widetilde{\mathbf{\Lambda}}_h^{\tilde{k}})^{-1} \phi(x,a)} \\
&\le \frac{\widetilde{\lambda}_y}{\sqrt{\widetilde{\lambda}_\Lambda}} \left\| \phi(x,a) \right\|_{(\widetilde{\mathbf{\Lambda}}_h^{\tilde{k}})^{-1}}
\end{aligned}$$

Finally, since we have $\langle \phi(x,a), \boldsymbol{w}_h^{\tilde{k}} \rangle - Q_h^\pi(x,a) = \langle \phi(x,a), \boldsymbol{w}_h^{\tilde{k}} - \boldsymbol{w}_h^\pi \rangle = \langle \phi(x,a), \mathbf{q}_1 + \mathbf{q}_2 + \mathbf{q}_3 + \mathbf{q}_4 \rangle$, we can write:

$$\begin{aligned}
&\left| \langle \phi(x,a), \boldsymbol{w}_h^{\tilde{k}} \rangle - Q_h^\pi(x,a) - \mathbb{P}_h(V_{h+1}^{\tilde{k}} - V_{h+1}^\pi)(x,a) \right| \\
&\leq \left( 2H\sqrt{d\widetilde{\lambda}_\Lambda} + 6dH\sqrt{\log(\chi)} + 2H\sqrt{d\widetilde{\lambda}_\Lambda} + \frac{\widetilde{\lambda}_y}{\sqrt{\widetilde{\lambda}_\Lambda}} \right) \|\phi(x,a)\|_{(\widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}})^{-1}}
\end{aligned}$$

Hence, for the inequality to hold, we need to set

$$\beta \geq 2H\sqrt{d\widetilde{\lambda}_\Lambda} + 6dH\sqrt{\log(\chi)} + 2H\sqrt{d\widetilde{\lambda}_\Lambda} + \frac{\widetilde{\lambda}_y}{\sqrt{\widetilde{\lambda}_\Lambda}} \tag{27}$$

Observe that the RHS of inequality (27) becomes:

$$\begin{aligned}
\text{RHS of Equation (27)} &= 2H\sqrt{d\widetilde{\lambda}_\Lambda} + 6dH\sqrt{\log(\chi)} + 2H\sqrt{d\widetilde{\lambda}_\Lambda} + \frac{\widetilde{\lambda}_y}{\sqrt{\widetilde{\lambda}_\Lambda}} \\
&\leq 4H\sqrt{d\widetilde{\lambda}_\Lambda} + 6dH\sqrt{\log(\chi)} + \frac{\widetilde{\lambda}_y}{\sqrt{\widetilde{\lambda}_\Lambda}}
\end{aligned}$$

Finally, we upper bound the ratio $\left(\widetilde{\lambda}_y\right) / \sqrt{\widetilde{\lambda}_\Lambda}$. Recall that $\widetilde{\lambda}_y = \sqrt{\frac{\left(H \cdot \frac{dH}{\log 2} \log\left(1 + \frac{K}{\widetilde{\lambda}_\Lambda d}\right)\right)^2 \log\left(\frac{3KH}{p}\right)}{\rho}}$ and $\widetilde{\lambda}_\Lambda = \frac{\log(K)\left(6\sqrt{d+1} + 2\log\left(\frac{3KH}{p}\right)\right)}{\sqrt{2\rho}}$, therefore

$$\frac{\widetilde{\lambda}_y}{\sqrt{\widetilde{\lambda}_\Lambda}} = \frac{\sqrt{2}\left(\frac{dH^2}{\log 2} \log\left(1 + \frac{K}{\widetilde{\lambda}_\Lambda d}\right)\right)\sqrt{\log\left(3KH/p\right)}}{\sqrt{\log\left(K\right)\left(6\sqrt{d+1} + \log\left(3KH/p\right)\right)}}$$

Observe that we can upper bound the last term by $H^2\sqrt{d\widetilde{\lambda}_\Lambda \log(\chi)}$. Therefore, it suffices to set $\beta = 5H^2\sqrt{d\widetilde{\lambda}_\Lambda \log(\chi)} + 6dH\sqrt{\log(\chi)}$.

**Proof of Theorem 27 (Lemma B.5 in Jin et al. (2020))** We will prove this lemma by induction. Base case: At the last step $H$, the statement is true because $Q_H^k(x,a) \geq Q_H^*(x,a)$. Since the value function at step $H+1$ is zero, by Theorem 26, we have:

$$|\langle \phi(x,a), \boldsymbol{w}_H^{\tilde{k}} \rangle - Q_H^*(x,a)| \leq \beta\sqrt{\phi(x,a)^\top (\widetilde{\boldsymbol{\Lambda}}_H^{\tilde{k}})^{-1} \phi(x,a)}$$

Hence, we have:

$$Q_H^*(x,a) \leq \min\{\langle \phi(x,a), \boldsymbol{w}_H^{\tilde{k}} \rangle + \beta\sqrt{\phi(x,a)^\top (\widetilde{\boldsymbol{\Lambda}}_H^{\tilde{k}})^{-1} \phi(x,a)}, H\} = Q_H^{\tilde{k}}(x,a).$$

Induction hypothesis: Suppose the statement is true at step $h+1$. Consider step $h$. By Theorem 26, we have:

$$|\langle \phi(x,a), \boldsymbol{w}_h^{\tilde{k}} \rangle| - Q_h^*(x,a) - \mathbb{P}_h(V_{h+1}^{\tilde{k}} - V_{h+1}^*)(x,a) \leq \beta\sqrt{\phi(x,a)^\top (\widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}})^{-1} \phi(x,a)}$$

By the induction assumption that $\mathbb{P}_h(V_{h+1}^{\tilde{k}} - V_{h+1}^*)(x,a) \geq 0$. we have:

$$Q_h^*(x.a) \leq \min\{\langle \phi(x,a), \boldsymbol{w}_h^{\tilde{k}} \rangle + \beta\sqrt{\phi(x,a)^\top (\widetilde{\boldsymbol{\Lambda}}_h^{\tilde{k}})^{-1} \phi(x,a)}, H\} = Q_h^k(x,a)$$

This concludes the proof.

## C. Switching Cost Analysis

First, we give an upper bound on the determinant of $\widetilde{\mathbf{\Lambda}}_h^k$.

**Lemma 29** (Similar to Lemma C.1 in (Wang et al., 2021)). *Let $\{\widetilde{\mathbf{\Lambda}}_h^k, (k,h) \in [K] \times [H]\}$ be as defined in Algorithm 1. Then for all $h \in [H]$ and $k \in [K]$, we have $\det\left(\widetilde{\mathbf{\Lambda}}_h^k\right) \leq (\widetilde{\lambda}_\Lambda + (k-1)/d)^d$*

*Proof.* We have

$$\begin{aligned}\mathrm{Tr}(\widetilde{\mathbf{\Lambda}}_h^k) &= \mathrm{Tr}(2\widetilde{\lambda}\boldsymbol{I} + \boldsymbol{M}_h^k) + \sum_{i=1}^{k-1} \mathrm{Tr}(\phi(x_h^i, a_h^i)\phi(x_h^i, a_h^i)\top) \\ &= \widetilde{\lambda}_\Lambda d + \sum_{i=1}^{k-1} \left\|\phi(x_h^i, a_h^i)\right\|_2^2 \leq \widetilde{\lambda}_\Lambda d + k - 1\end{aligned}$$

where the last inequality is because we assume $\|\phi(x,a)\| \leq 1$. Since $\widetilde{\mathbf{\Lambda}}_h^k$ is PSD, by AM-GM, we have

$$\det(\widetilde{\mathbf{\Lambda}}_h^k) \leq \left(\frac{\mathrm{Tr}(\widetilde{\mathbf{\Lambda}}_h^k)}{d}\right)^d \leq \left(\widetilde{\lambda}_\Lambda + \frac{k-1}{d}\right)^d$$

□

Next, we provides a determinant-based upper bound for the ratio between the norms $\|\cdot\|_A$ and $\|\cdot\|_B$, where $A \succeq B$.

**Lemma 30** (Lemma 12 in (Abbasi-Yadkori et al., 2011)). *Suppose $A, B \in \mathbb{R}^{d\times d}$ are two PSD matrices such that $A \succeq B$, then for any $x \in \mathbb{R}^d$, we have $\|x\|_A \leq \|x\|_B \cdot \sqrt{\det(A)/\det(B)}$*

Finally, we can derive the switching cost of Algorithm 1 in the following lemma:

**Lemma 31** (Similar to Lemma C.3 in (Wang et al., 2021)). *Condition on the event that $\left\|\boldsymbol{M}_h^k\right\| \leq \widetilde{\lambda}_\Lambda$ for all $h, k \in [H]\times[K]$. For $C = 2$ and $\widetilde{\lambda}_\Lambda > 0$, the global switching cost of Algorithm 1 is bounded by:*

$$N_{\mathit{count}} \leq \frac{dH}{\log 2} \log\left(1 + \frac{K}{\widetilde{\lambda}_\Lambda d}\right)$$

*Proof.* Let $\{k_1, k_2, \cdots, k_{N_{\text{count}}}\}$ be the episodes where Algorithm 1 updates the policy, and let $k_0 = 0$. Then, by the update condition on line (8), for each $i \in [N_{\text{count}}]$, there exists at least one $h \in [H]$ such that

$$\det(\widetilde{\mathbf{\Lambda}}_h^{k_i}) > 2\det(\widetilde{\mathbf{\Lambda}}_h^{k_{i-1}})$$

By the definition of $\widetilde{\mathbf{\Lambda}}_h^k$, we know that $\widetilde{\mathbf{\Lambda}}_h^{i_1} \succeq \widetilde{\mathbf{\Lambda}}_h^{i_2}$ for all $i_1 \geq i_2$ and $h \in [H]$. Hence, we have

$$\prod_{h=1}^H \det(\widetilde{\mathbf{\Lambda}}_h^{k_i}) > 2\prod_{h=1}^H \det(\widetilde{\mathbf{\Lambda}}_h^{k_{i-1}})$$

We can recursively apply the inequality above to all $i \in [N_{\text{count}}]$ and get

$$\prod_{h=1}^H \det(\widetilde{\mathbf{\Lambda}}_h^{k_{N_{\text{count}}}}) > 2^{N_{\text{count}}} \cdot \prod_{h=1}^H \det(\widetilde{\mathbf{\Lambda}}_h^1) = 2^{N_{\text{count}}}\widetilde{\lambda}_\Lambda^{dH}$$

as we initialize $\widetilde{\mathbf{\Lambda}}_h^1 = 2\widetilde{\lambda}_\Lambda \boldsymbol{I}$.

Also, by Theorem 29, we have

$$\prod_{h=1}^{H} \det\left(\widetilde{\mathbf{\Lambda}}_h^{k_{N_{\text{count}}}}\right) \leq \prod_{h=1}^{H} \det(\widetilde{\mathbf{\Lambda}}_h^{K}) \leq \left(\widetilde{\lambda}_\Lambda + \frac{K}{d}\right)^{dH}$$

Therefore, we can combine the two inequalities above and get

$$N_{\text{count}} \leq \frac{dH}{\log 2} \log\left(1 + \frac{K}{\widetilde{\lambda}_\Lambda d}\right)$$

□

## D. Auxiliary Results

**Claim 32** (Concentration inequalities in (Tao, 2012)). *Let $M \in \mathbb{R}^{d\times d}$ be a symmetric matrix where each of its entries $M_{i,j} = M_{j,u} \sim \mathcal{N}(0,1)$ for any $1 \leq i \leq j \leq d$. Then, for any $\alpha > 0$, $\mathbb{P}(\|M\|_{op} \geq 4\sqrt{d} + 2\log\left(\frac{1}{\alpha}\right)) \leq \alpha$, where $\|M\|_{op}$ is the operator norm of a matrix associated to the norm $\|\cdot\|_2$.*

**Claim 33** (Corollary to Lemma 1 in (Laurent, 2005)). *If $U \sim \chi^2(d)$ and $\alpha \in (0,1)$:*

$$\mathbb{P}\left(U \geq d + 2\sqrt{d\log\left(\frac{1}{\alpha}\right) + 2\log\left(\frac{1}{\alpha}\right)}\right) \leq \alpha$$

$$\mathbb{P}\left(U \leq d - 2\sqrt{d\log\left(\frac{1}{\alpha}\right)}\right) \leq \alpha$$

*As a consequence of the first inequality, we also have that for any vector $v \in \mathbb{R}^d$ drawn from a $d-$dimensional Gaussian distribution $\mathcal{N}(0, \boldsymbol{I}_{d\times d})$, then $\mathbb{P}\left(\|v\|_2 > \sqrt{d} + 2\sqrt{\log\left(\frac{1}{\alpha}\right)}\right) \leq \alpha$.*

*Proof.* By definition of Laplace distribution, for each $i \in [d]$, we have

$$\mathbb{P}\left(|v_i| > \log\left(\frac{d}{\alpha}\right)\right) \leq \frac{\alpha}{d}$$

Hence, in the union event, with probability at least $1-\alpha$, we have $\|v\|_2 \leq \sqrt{d}\log\left(\frac{d}{\alpha}\right)$. □

**Claim 34** (Theorem 7.8 in (Zhang, 2011)). *For two positive definite matrices (PSD) $A, B \in \mathbb{R}^{d\times d}$, we write $A \succeq B$ to denote that $A - B$ is PSD. Then, if $A \succeq B \succeq 0$, we have:*

- $\text{rank}(A) \geq \text{rank}(B)$
- $\det(A) \geq \det(B)$
- $B^{-1} \succeq A^{-1}$ *if $A$ and $B$ are non-singular.*

**Claim 35** (Lemma 12 in (Abbasi-Yadkori et al., 2011)). *Supposed $A, B \in \mathbb{R}^{d\times d}$ are two PSD matrices such that $A \succeq B$. Then, for any $x \in \mathbb{R}^d$, we have $\|x\|_A \leq \|x\|_B \sqrt{\frac{\det(A)}{\det(B)}}$.*