Frequentist Regret Bounds for Randomized Least-Squares Value Iteration

Andrea Zanette*
Stanford University

David Brandfonbrener*
New York University

Emma Brunskill
Stanford University

Matteo Pirotta
Facebook AI Research

Alessandro Lazaric
Facebook AI Research

Abstract

We consider the exploration-exploitation dilemma in finite-horizon reinforcement learning (RL). When the state space is large or continuous, traditional tabular approaches are unfeasible and some form of function approximation is mandatory. In this paper, we introduce an optimistically-initialized variant of the popular randomized least-squares value iteration (RLSVI), a model-free algorithm where exploration is induced by perturbing the least-squares approximation of the action-value function. Under the assumption that the Markov decision process has low-rank transition dynamics, we prove that the frequentist regret of RLSVI is upper-bounded by $\widetilde{O}(d^2 H^2 \sqrt{T})$ where d is the feature dimension, H is the horizon, and T is the total number of steps. To the best of our knowledge, this is the first frequentist regret analysis for randomized exploration with function approximation.

1 Introduction

A key challenge in reinforcement learning (RL) is how to balance exploration and exploitation in order to efficiently learn to make good sequences of decisions in a way that is both computationally tractable and statistically efficient. In the tabular case, the exploration-exploitation problem is well-understood for a number

*Equal Contribution

of settings (e.g., finite-horizon, average reward, infinite horizon with discount), exploration objectives (e.g., regret minimization and probably approximately correct), and for different algorithmic approaches, where optimism-under-uncertainty (Jaksch et al., 2010; Fruit et al., 2018) and Thompson sampling (TS) (Osband et al., 2016a; Russo, 2019) are the most popular principles. For instance, in the finite-horizon setting, Azar et al. (2017) and Zanette and Brunskill (2019) recently derived minimax optimal and structure adaptive regret bounds for optimistic exploration algorithms. TS-based algorithms have mainly been analyzed in tabular MDPs in terms of Bayesian regret (Osband et al., 2016a; Osband and Roy, 2017; Ouyang et al., 2017), which assumes that the MDP is sampled from a known prior distribution. These bounds do not hold against a fixed MDP and algorithms with small Bayesian regret may still suffer high regret in some hard-to-learn MDPs within the chosen prior. In the tabular setting, frequentist (or worst-case) regret analysis has been developed for TS-based algorithms both in the average reward (Gopalan and Mannor, 2015; Agrawal and Jia, 2017) and finite-horizon case (Russo, 2019). Despite the fact that TS-based approaches have slightly worse regret bounds compared to optimism-based algorithms, their empirical performance is often superior (Chapelle and Li, 2011; Osband and Roy, 2017).

Unfortunately, the performance of tabular exploration methods rapidly degrades with the number of states and actions, thus making them infeasible in large or continuous MDPs. So, one of the most important challenges to improve sample efficiency in large-scale RL is how to combine exploration mechanisms with generalization methods to obtain algorithms with provable regret guarantees. The simplest approach to deal with continuous state is discretization. It has been used in Ortner and Ryabko (2012); Lakshmanan et al. (2015) to derive $\widetilde{O}(T^{3/4})$ and $\widetilde{O}(T^{2/3})$ frequentistic regret bounds for average reward MDPs. Recent work

on contextual MDPs (Jiang et al., 2017; Dann et al., 2018) yielded promising sample efficiency guarantees, but such algorithms are computationally intractable, and their bounds are not tight in the tabular settings.

One of the most simple and popular forms of function approximation is to use a linear representation for the action-value functions. When the transition model also has low-rank structure, very recent work has shown that a variant of Q-learning can achieve polynomial sample complexity as a function of the state space dimension when given access to a generative model (Yang and Wang, 2019b). Nonetheless, the generative model assumption removes most of the exploration challenge, as the state space can be arbitrarily sampled. Concurrently to our work, optimism-based exploration has been successfully integrated with linear function approximation both in model-based and model-free algorithms (Yang and Wang, 2019a; Jin et al., 2019). In MDPs with low-rank dynamics, these algorithms are proved to have regret bounds scaling with the dimensionality d of the linear space (i.e., the number of features) instead of the number of states.

On the algorithmic side, TS-based exploration can be easily integrated with linear function approximation as suggested in the Randomized Least-Squares Value Iteration (RLSVI) algorithm (Osband et al., 2016b). Despite promising empirical results, RLSVI has been analyzed only in the tabular case (i.e., when the features are indicators for each state) and for Bayesian regret. While RLSVI is a model-free algorithm, recent work (Russo, 2019) leverages an equivalence between model-free and model-based algorithms in the tabular case to derive frequentist regret bounds. The analysis carefully chooses the variance of the perturbations applied to the estimated solution to ensure that the value estimates are optimistic with constant probability.

In this paper we provide the first frequentist regret analysis for a variant of RLSVI when linear function approximation is used in the finite-horizon setting. Similar to optimistic PSRL for the tabular setting (Agrawal and Jia, 2017), we modify RLSVI to ensure that the perturbed estimates used in the value iteration process are optimistic with constant probability. Following the results in the linear bandit literature (Abeille et al., 2017), we show that the perturbation applied to the the least-squares estimates should be larger than their estimation error. However, in contrast to bandit, perturbed estimates are propagated back through iterations and we need to carefully adjust the perturbation scheme so that the probability of being optimistic does not decay too fast with the horizon and, at the same time, we can control how the perturbations accumulate over iterations. Under the assumption that the system dynamics are low-rank,

we show that the frequentist regret of our algorithm is $\widetilde{O}(H^2 d^2 \sqrt{T} + H^5 d^4 + \epsilon dH(1 + \epsilon dH^2)T)$ where ϵ is the misspecification level, H is the fixed horizon, d is the number of features, and T is the number of samples. Similar to linear bandits, this is worse by a factor of $\sqrt{Hd}$ (i.e., the square root of the dimension of the estimated parameters) than the optimistic algorithm of Jin et al. (2019). Whether this gap can be closed is an open question both in bandits and RL.

2 Preliminaries

We consider an undiscounted finite-horizon MDP (Puterman, 1994) $M = (\mathcal{S}, \mathcal{A}, \mathbb{P}, r, H)$ with state space $\mathcal{S}$, action space $\mathcal{A}$ and horizon length $H \in \mathbb{N}^+$. For every $t \in [H] \stackrel{def}{=} \{1, \ldots, H\}$, every state-action pair is characterized by a reward $r_t(s, a) \in [0, 1]$ and a transition kernel $\mathbb{P}_t(\cdot|s, a)$ over next state. We assume $\mathcal{S}$ to be a measurable, possibly infinite, space and $\mathcal{A}$ can be any (compact) time and state dependent set (we omit this dependency for brevity). For any $t \in [H]$ and $(s, a) \in \mathcal{S} \times \mathcal{A}$, the state-action value function of a non-stationary policy $\pi = (\pi_1, \ldots, \pi_H)$ is defined as $Q_t^\pi(s, a) = r_t(s, a) + \mathbb{E}\left[\sum_{l=t+1}^{H} r_l(s_l, \pi_l(s_l)) \mid s, a\right]$ and the value function is $V_t^\pi(s) = Q_t^\pi(s, \pi_t(s))$. Since the horizon is finite, under some regularity conditions, (Shreve and Bertsekas, 1978), there always exists an optimal policy $\pi^\star$ whose value and action-value functions are defined as $V_t^\star(s) \stackrel{def}{=} V_t^{\pi^\star}(s) = \sup_\pi V_t^\pi(s)$ and $Q_t^\star(s, a) \stackrel{def}{=} Q_t^{\pi^\star}(s, a) = \sup_\pi Q_t^\pi(s, a)$. Both Q^π and $Q^\star$ can be conveniently written as the result of the Bellman equations

$$Q_t^\pi(s, a) = r_t(s, a) + \mathbb{E}_{s' \sim \mathbb{P}_t(\cdot|s,a)}[V_{t+1}^\pi(s')] \quad (1)$$

$$Q_t^\star(s, a) = r_t(s, a) + \mathbb{E}_{s' \sim \mathbb{P}_t(\cdot|s,a)}[V_{t+1}^\star(s')] \quad (2)$$

where $V_{H+1}^\pi(s) = V_{H+1}^\star(s) = 0$ and $V_t^\star(s) = \max_{a \in \mathcal{A}} Q_t^\star(s, a)$, for all $s \in \mathcal{S}$. Notice that by boundedness of the reward, for any t and (s, a), all functions $Q_t^\pi, V_t^\pi, Q_t^\star, V_t^\star$ are bounded in $[0, H - t + 1]$.

The learning problem The learning agent interacts with the MDP in a sequence of episodes $k \in [K]$ of fixed length H by playing a nonstationary policy $\pi_k = (\pi_{1k}, \ldots, \pi_{Hk})$ where $\pi_{tk} : \mathcal{S} \to \mathcal{A}$. In each episode, the initial state s_{1k} is chosen arbitrarily and revealed to the agent. The learning agent does not know the transition or reward functions, and it relies on the samples (i.e., states and rewards) observed over episodes to improve its performance over time. Finally, we evaluate the performance of an agent by its regret after K episodes: $\text{REGRET}(K) \stackrel{def}{=} \sum_{k=1}^{K} V_1^\star(s_{1k}) - V_1^{\pi_k}(s_{1k})$.

Linear function approximation and low-rank MDPs. Whenever the state space $\mathcal{S}$ is too large

or continuous, functions above cannot be represented by enumerating their values at each state or state-action pair. A common approach is to define a feature map $\phi_t : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, possibly different at any $t \in [H]$, embedding each state-action pair (s, a) into a d-dimensional vector $\phi_t(s, a)$. The action-value functions are then represented as a linear combination between the features ϕ_t and a vector parameter $\theta_t \in \mathbb{R}^d$, such that $Q_t(s, a) = \phi_t(s, a)^\top \theta_t$. This representation effectively reduces the complexity of the problem from $\mathcal{S} \times \mathcal{A}$ down to d. Nonetheless, $Q_t^\star$ may not fit into the space spanned by ϕ_t, and approximate value iteration may propagate and accumulate errors over iterations (Munos, 2005; Munos and Szepesvári, 2008), and an exploration algorithm may suffer linear regret. Thus, similar to (Yang and Wang, 2019a,b; Jin et al., 2019), we consider MPDs that are "coherent" with the feature map ϕ_t used to represent action-value functions. In particular, we assume that M has (approximately) low-rank transition dynamics and linear reward in ϕ_t.

Assumption 1 (Approximately Low-Rank MDPs). *We assume that for each $t \in [H]$ there exist a feature map $\psi_t : \mathcal{S} \to \mathbb{R}^d$, $s \mapsto \psi_t(s)$ and a parameter $\theta_t^r \in \mathbb{R}^d$ such that the reward can be decomposed as a linear response and a non-linear term:*

$$r_t(s, a) = \phi_t(s, a)^\top \theta_t^r + \Delta_t^r(s, a) \qquad (3)$$

and the dynamics are approximately low-rank:

$$\mathbb{P}_t(s' \mid s, a) = \phi_t(s, a)^\top \psi_t(s') + \Delta_t^P(s' \mid s, a). \qquad (4)$$

We denote by ϵ an upper bound on the non-linear terms, as follows:

$$|\Delta_t^r(s, a)| \leqslant \epsilon, \qquad \|\Delta_t^P(\cdot \mid s, a)\|_1 \leqslant \epsilon. \qquad (5)$$

We further make the following regularity assumptions:

$$\|\phi_t(s, a)\|_2 \leqslant L_\phi, \; \|\theta_t^r\|_2 \leqslant L_r, \; \int_s \|\psi_t(s)\| \leqslant L_\psi. \quad (6)$$

An important consequence of Asm. 1 in the absence of misspecification ($\epsilon = 0$) is that the Q-function of any policy is linear in the features ϕ.

Proposition 1. *If $\epsilon = 0$, for every policy π and timestep $t \in [H]$ there exists $\theta_t^\pi \in \mathbb{R}^d$ such that*

$$Q_t^\pi(s, a) = \phi_t(s, a)^\top \theta_t^\pi, \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A}. \quad (7)$$

Proof. The definition of low-rank MDP from Asm. 1 together with the Bellman equation gives:

$$\begin{aligned} Q_t^\pi(s, a) &= r_t(s, a) + \mathbb{E}_{s'|s,a}[V_{t+1}^\pi(s')] \\ &= \phi_t(s, a)^\top \theta_t^r + \int_{s'} \phi_t(s, a)^\top \psi_t(s') V_{t+1}^\pi(s') \\ &= \phi_t(s, a)^\top \left(\theta_t^r + \int_{s'} \psi_t(s') V_{t+1}^\pi(s') \right) \quad (8) \end{aligned}$$

We define θ_t^π to be the term inside the parentheses. $\square$

To give further intuition about the assumption, consider the case of finite state and action spaces (again with $\epsilon = 0$). Then we can write:

$$\mathbb{P}_t(s, a) = \phi_t(s, a)^\top \Psi_t \qquad (9)$$

for a certain $\Psi_t \in \mathbb{R}^{d \times \mathcal{S}}$. Then for any policy π there exists a matrix Φ^π such that the transition matrix of the Markov chain P^π can be expressed by a low-rank factorization:

$$P_t^\pi = \Phi_t^\pi \Psi_t, \quad \Phi_t^\pi \in \mathbb{R}^{S \times d}, \Psi_t \in \mathbb{R}^{d \times S} \qquad (10)$$

where in particular Φ_t^π depends on the policy π:

$$\Phi_t^\pi[s, :] = \phi_t(s, \pi(s))^\top, \quad \Psi_t[:, s'] = \psi_t(s'). \qquad (11)$$

Since $\text{RANK}(\Phi_t^\pi) \leqslant d, \text{RANK}(\Psi_t) \leqslant d$ we get $\text{RANK}(P_t^\pi) \leqslant d$ (see Golub and Van Loan (2012)).

3 Algorithm

Our primary goal in this work is to provide a Thompson sampling (TS)-based algorithm with linear value function approximation with frequentist regret bounds. A key challenge in frequentist analyses of TS algorithms is to ensure sufficient exploration using randomized (i.e., perturbed) versions of the estimated model or value function. A common way to obtain effective exploration has been to consider perturbations large enough so that the resulting sampled model or value function is optimistic with a fixed probability (Agrawal and Goyal, 2013; Abeille et al., 2017; Russo, 2019). However, such prior work has only considered the bandit or tabular MDP settings. Here we modify RLSVI described by Osband et al. (2016b) to use an optimistic "default" value function during an initial phase and inject carefully-tuned perturbations to enable *frequentist* regret bounds in low-rank MDPs. We refer to the resulting algorithm as OPT-RLSVI and we illustrate it in Alg. 1.

Gaussian noise to encourage exploration. OPT-RLSVI proceeds in episodes. At the beginning of episode k it receives an initial state s_{1k} and runs a value iteration procedure to compute a linear approximation of $Q_t^\star$ at each timestep $t \in [H]$. To encourage exploration, the learned parameter $\widehat{\theta}_{tk}$ is perturbed by adding mean-zero Gaussian noise $\overline{\xi}_{tk} \sim \mathcal{N}(0, \sigma^2 \Sigma_{tk}^{-1})$, obtaining $\overline{\theta}_{tk} = \widehat{\theta}_{tk} + \overline{\xi}_{tk}$. The perturbation (or *pseudonoise*) $\overline{\xi}_{tk}$ has variance proportional to the inverse of the regularized design matrix $\Sigma_{tk} = \sum_{i=1}^{k-1} \phi_{ti} \phi_{ti}^\top + \lambda I$, where the ϕ_{ti}'s are the features encountered in prior episodes; this results in perturbations with higher variance in less explored directions. Finally, we show how to choose the magnitude σ^2 of the variance in Sec. 5.2 to ensure sufficient exploration.

A key contribution of our work is to prove that this strategy can guarantee reliable exploration under Asm. 1. We do this by showing that the algorithm is optimistic with constant probability. Explicitly, we prove that the (random) value function difference $(\overline{V}_{1k} - V_1^\star)(s_{1k})$ can be expressed as a *one-dimensional biased random walk*, which depends on a high probability bound on the environment noise (the bias of the walk) and on the variance of the injected pseudonoise (the variance of the walk). By setting the pseudonoise to have the appropriate variance we can guarantee that the random walk is "optimistic" enough that the algorithm explores sufficiently. Unfortunately, it is possible to analyze the algorithm as a random walk only if the value function is not perturbed by clipping; otherwise, one cannot write down the walk and the process is difficult to analyze as further bias is introduced by clipping. However, not clipping the value function may give rise to abnormal values.

The issue of abnormal values. A common problem that arises in estimation in RL with function approximation is that as a result of statistical errors combined with the bootstrapping and extrapolation of the next-state value function (Munos, 2005; Munos and Szepesvári, 2008; Farahmand et al., 2010) the value function estimate can take values outside its plausible range. A common solution is to "clip" the bootstrapped value function into the range of plausible values (in this case, between 0 and H). This avoids propagating overly abnormal values to the estimated parameters at prior timesteps which would degrade their estimation accuracy. Clipping the value function is also a solution typically employed in tabular algorithms for exploration (Azar et al., 2017; Dann et al., 2017; Zanette and Brunskill, 2019; Yang and Wang, 2019a; Dann et al., 2019). After adding optimistic bonuses for exploration they "clip" the value function above by H, which is an upper bound on the true optimal value function. Since H is guaranteed to be an optimistic estimate for $V^\star$, clipping effectively preserves optimism while keeping the value function bounded for bootstrapping. However, clipping cannot be easily integrated in our setting as it effectively introduces bias in the pseudonoise and it may "pessimistically" affect the value function estimates, reducing the probability of being optimistic.

Default value function. To avoid propagating unreasonable values without using clipping, we define a default value function, similar in the spirit to algorithms such as $R_{\max}$ (Brafman and Tennenholtz, 2002). In particular, we assign the maximum plausible value $\overline{Q}_t(s, a) = H - t + 1$ to an uncertain direction $\phi_t(s, a)$ (as measured by the $\|\phi_t(s, a)\|_{\Sigma_{tk}^{-1}}$ norm). Once a given direction $\phi_t(s, a)$ has been tried a suf-

ficient number of times we can guarantee (under an inductive argument) that the linearity of the representation is accurate enough that with high probability $\phi_t(s, a)^\top \overline{\theta}_{tk} - Q_t^\star(s, a) \in [-(H - t + 1), 2(H - t + 1)]$. In other words, abnormal values are not going to be encountered, and thus clipping becomes unnecessary. Notice that this accuracy requirement is quite minimal because $V_t^\star$ has a range of at most $H - t + 1$.

We emphasize that the purpose of the optimistic default function is not to inject further optimism but rather to keep the propagation of the errors under control while ensuring optimism.

Defining the $\overline{Q}$ values. Finally, we also choose our Q function to interpolate between the "default" optimistic value and the linear function of the features as the uncertainty decreases. The main reason is to ensure *continuity* of the function, which facilitates the handling of some of the technical aspects connected to the concentration inequality (in particular in App. E).

Definition 1 (Algorithm Q function). *For some constants α_L, α_U and using shorthand for the feature $\phi \overset{def}{=} \phi_t(s, a)$, the default function $B_t \overset{def}{=} H - t + 1$ and the interpolation parameter $\rho \overset{def}{=} \frac{\|\phi\|_{\Sigma_{tk}^{-1}} - \alpha_L}{\alpha_U - \alpha_L}$ define:*

$$\overline{Q}_{tk}(s, a) \overset{def}{=} \begin{cases} \phi^\top \overline{\theta}_{tk}, & \text{if } \|\phi\|_{\Sigma_{tk}^{-1}} \leqslant \alpha_L \\ B_t, & \text{if } \|\phi\|_{\Sigma_{tk}^{-1}} \geqslant \alpha_U \\ \rho\left(\phi^\top \overline{\theta}_{tk}\right) + (1 - \rho)B_t, & \text{otherwise.} \end{cases}$$

Algorithm 1 OPT-RLSVI

1: Initialize $\Sigma_{t1} = \lambda I$, $\forall t \in [H]$; Define $\overline{V}_{tk}(s) = \max_a \overline{Q}_{tk}(s, a)$, with $\overline{Q}_{tk}(s, a)$ defined in Def. 1
2: **for** $k = 1, 2, \ldots$ **do**
3:     Receive starting state s_{1k}
4:     Set $\overline{\theta}_{H+1,k} = 0$
5:     **for** $t = H, H - 1, \ldots, 1$ **do**
6:       $\widehat{\theta}_{tk} = \Sigma_{tk}^{-1}\left(\sum_{i=1}^{k-1} \phi_{ti}[r_{ti} + \overline{V}_{t+1,k}(s_{t+1,i})]\right)$
7:       Sample $\overline{\xi}_{tk} \sim \mathcal{N}(0, \sigma^2 \Sigma_{tk}^{-1})$
8:       $\overline{\theta}_{tk} = \widehat{\theta}_{tk} + \overline{\xi}_{tk}$
9:     **end for**
10:    Execute $\overline{\pi}_{tk}(s) = \arg\max_a \overline{Q}_{tk}(s, a)$, see Def. 1
11:    Collect trajectories of (s_{tk}, a_{tk}, r_{tk}) for $t \in [H]$.
12:    Update $\Sigma_{t,k+1} = \Sigma_{tk} + \phi_{tk}\phi_{tk}^\top$ for $t \in [H]$
13: **end for**

---

4 Main Result

We present the first frequentist regret bound for a TS-based algorithm in MDPs with approximate linear reward response and low-rank transition dynamics:

Theorem 1. *Assume Asm. 1 and set the algorithm parameters* $\lambda = 1$, $\sigma = \sqrt{H\nu_k(\delta)} = \sqrt{H}(\tilde{O}(Hd) + L_\phi(3HL_\psi + L_r) + 4\epsilon H\sqrt{dk})$, $\alpha_U = 1/\tilde{O}(\sigma\sqrt{d})$, *and* $\alpha_L = \alpha_U/2$ *(full definitions with the log terms can be found in App. D). Then for any* $0 < \delta < \Phi(-1)/2$, *with probability at least* $1 - \delta$ *the regret of* OPT-RLSVI *is bounded jointly for all episodes* k *up to* K *by:*

$$\tilde{O}\left(\sigma dH\sqrt{K} + \frac{H^2 d}{\alpha_L^2} + \epsilon H^2 K\right). \qquad (12)$$

If we further assume that $L_\phi = \tilde{O}(1)$ *and* $L_r, L_\psi = \tilde{O}(d)$, *then the bound reduces to*

$$\tilde{O}\left(H^2 d^2\sqrt{T} + H^5 d^4 + \epsilon dH(1 + \epsilon dH^2)T\right). \qquad (13)$$

For the setting of low-rank MDPs a lower bound is currently missing both in terms of statistical rate and regarding the misspecification. Recently, Du et al. (2019); Lattimore and Szepesvari (2019); Van Roy and Dong (2019) discuss what's possible to achieve regarding the misspecification level.

For finite action spaces OPT-RLSVI can be implemented efficiently in space $O(d^2 H + dAHK)$ and time $O(d^2 AHK^2)$ where A is the number of actions (Prop. H.1 in appendix).

It is useful to compare our result with Yang and Wang (2019a) and Jin et al. (2019) which study a similar setting but with an approach based on deterministic optimism, and with Russo (2019) which proves worst-case regret bounds of RLSVI for tabular representations.

Comparison with Yang and Wang (2019a). Recently, Yang and Wang (2019a) studied exploration in finite state-spaces and low-rank transitions. They define a model-based algorithm that tries to learn the "core matrix", defined as the middle factor of a three-factor low-rank factorization. While their regularity assumptions on the parameters do not immediately fit in our framework, an important distinction (beyond model-based vs model-free) is that their algorithm potentially needs to compute the value function across all states. This suffers $\Omega(S)$ computational complexity and cannot directly handle continuous state spaces.

Comparison with Jin et al. (2019). A more direct comparison can be done with Jin et al. (2019) which is based on least-square value iteration (like OPT-RLSVI) and uses the same setting as we do when $L_r = L_\psi = \sqrt{d}$ and $L_\phi = 1$. In that case we get the regret in Eqn. (13) which is $\sqrt{Hd}$-times worse in the leading term than Jin et al. (2019).

In terms of feature dimension d, this matches the $\sqrt{d}$ gap in linear bandits between the best bounds for a TS-based algorithm (with regret $\tilde{O}(d^{3/2}\sqrt{T})$) (Abeille et al., 2017) and the best bounds for an optimistic algorithm (with regret $\tilde{O}(d\sqrt{T})$) (Abbasi-Yadkori et al., 2011). This happens because the proof techniques for Thompson sampling require the perturbations to have sufficient variance to guarantee optimism (and thus exploration) with some probability. For a geometric interpretation of this, see Abeille et al. (2017). For H-horizon MDPs, the total system dimensionality is dH, and therefore the extra $\sqrt{dH}$ factor is expected.

Comparison with Russo (2019). Recently, Russo (2019) has analyzed RLSVI in tabular finite horizon MDPs. While the core algorithm is similar, function approximation does introduce challenges that required changing RLSVI by, e.g., introducing the default function. While in Russo (2019) the value function can be bounded in high probability thanks to the non-expansiveness of the Bellman operator associated to the estimated model, in our case this has to be handled explicitly. We think that the use of a default optimistic value function could yield better horizon dependence for RLSVI in tabular settings, though this would require changing the algorithm.

5 Proof Outline

In this section we outline the proof of our regret bound for OPT-RLSVI. The four main ingredients are: 1) a one-step expansion of the action-value function difference $\overline{Q}_{tk} - Q_t^{\pi_k}$ in terms of the next-state value function difference; 2) a high probability bound on the noise and pseudonoise; 3) showing that the algorithm is optimistic with constant probability; 4) combining to get the regret bound. For the sake of clarity, we will assume no misspecification ($\epsilon = 0$), no regularization ($\lambda = 0$), and a nonsingular design matrix $\Sigma_{tk} = \sum_{i=1}^{k-1} \phi_{ti}\phi_{ti}^\top$. The complete proof is reported in the appendix.

5.1 One-Step Analysis of Q functions

In this section we do a "one-step" analysis to decompose the difference in Q functions in the case where $\|\phi_t(s,a)\|_{\Sigma_{tk}^{-1}} \leq \alpha_L$ so that $\overline{Q}_{tk}$ is linear in the features. The decomposition has three parts: environment noise, pseudonoise, and the difference in value functions at step $t + 1$. It reads $(\overline{Q}_{tk} - Q_t^\pi)(s, a) =$

$$\phi_t(s,a)^\top(\overline{\eta}_{tk} + \overline{\xi}_{tk}) + \mathbb{E}_{s'|s,a}(\overline{V}_{t+1,k} - V_{t+1}^\pi)(s') \quad (14)$$

where $\overline{\eta}_{tk}$ is the projected environment noise defined below in Eqn. (18). The complete version of the decomposition is Lem. C.1 in the appendix, while here we give an informal proof sketch of this fact.

First, since we are assuming that $\|\phi_t(s,a)\|_{\Sigma_{tk}^{-1}} \leq \alpha_L$

and $\epsilon = 0$, we can apply Def. 1 and Prop. 1 to write:

$$(\overline{Q}_{tk} - Q_t^\pi)(s, a) = \phi_t(s, a)^\top (\overline{\theta}_{tk} - \theta_t^\pi). \qquad (15)$$

Decomposing $\overline{\theta}_{tk} = \widehat{\theta}_{tk} + \overline{\xi}_{tk}$ immediately shows how the pseudonoise $\overline{\xi}_{tk}$ appears in Eqn. (14). Now we need to handle the regression term:

$$\widehat{\theta}_{tk} \stackrel{def}{=} \Sigma_{tk}^{-1} \sum_{i=1}^{k-1} \phi_{ti}(r_{ti} + \overline{V}_{t+1,k}(s_{t+1,i})). \qquad (16)$$

To handle this, we need to make an expectation over s' given s_{ti}, a_{ti} (the experienced state and action in timestep t of episode i) appear in each term of the sum so that the value function term will become linear in ϕ_{ti}. To do this, we define the one-step environment noise with respect to $\overline{V}_{t+1,k}$ as

$$\overline{\eta}_{tk}(i) \stackrel{def}{=} \overline{V}_{t+1,k}(s_{t+1,i}) - \mathbb{E}_{s'|s_{ti},a_{ti}}[\overline{V}_{t+1,k}(s')], \quad (17)$$

Then we define the projected environment noise as:

$$\overline{\eta}_{tk} \stackrel{def}{=} \Sigma_{tk}^{-1} \sum_{i=1}^{k-1} \phi_{ti}\overline{\eta}_{tk}(i). \qquad (18)$$

Putting this into the definition of $\widehat{\theta}_{tk}$ from Eqn. (16),

$$\widehat{\theta}_{tk} = \Sigma_{tk}^{-1} \sum_{i=1}^{k-1} \phi_{ti}(r_{ti} + \mathbb{E}_{s'|s_{ti},a_{ti}}[\overline{V}_{t+1,k}(s')] + \overline{\eta}_{tk}(i))$$

$$= \overline{\eta}_{tk} + \Sigma_{tk}^{-1} \sum_{i=1}^{k-1} \phi_{ti}(r_{ti} + \mathbb{E}_{s'|s_{ti},a_{ti}}[\overline{V}_{t+1,k}(s')]).$$

But now we note that this reward plus expected value function is linear in the features (thanks to Prop. 1), so we can rewrite the second term as

$$\Sigma_{tk}^{-1} \sum_{i=1}^{k-1} \phi_{ti}\phi_{ti}^\top \left(\theta^r + \int_{s'} \psi_t(s')\overline{V}_{t+1,k}(s') \right) \qquad (19)$$

$$= \theta^r + \int_{s'} \psi_t(s')\overline{V}_{t+1,k}(s'). \qquad (20)$$

Finally, comparing with the definition of θ_t^π (Eqn. (8)) we see that the θ^r terms cancel and we get

$$\overline{\theta}_{tk} - \theta_t^\pi = \overline{\xi}_{tk} + \overline{\eta}_{tk} + \int_{s'} \psi_t(s')(\overline{V}_{t+1,k} - V_{t+1}^\pi)(s').$$

Premultiplying by $\phi_t(s, a)^\top$ gives Eqn. (14).

5.2 High probability bounds on the noise

To ensure that our estimates concentrate around the true Q functions, we need to ensure that the $\overline{\eta}_{tk}$ and $\overline{\xi}_{tk}$ are not too large. This is achieved with similar ideas of self-normalizing processes as is done for

linear bandits (Abbasi-Yadkori et al., 2011), with an additional union bound over possible value functions $\overline{V}_{t+1,k}$ which depend on $\overline{\theta}_{tk}$ and Σ_{tk}^{-1}. In the end, we prove in Lem. E.6 that indeed with high probability for any ϕ:

$$|\phi^\top \overline{\eta}_{tk}| \leqslant \|\phi\|_{\Sigma_{tk}^{-1}} \|\overline{\eta}_{tk}\|_{\Sigma_{tk}} = \sqrt{\nu_k(\delta)}\|\phi\|_{\Sigma_{tk}^{-1}} \qquad (21)$$

where $\sqrt{\nu_k(\delta)} = \widetilde{O}(dH)$ is defined fully in App. D. While we defer the computation of the "right" amount of pseudonoise to the next subsection, here we mention that for the choice we make $\overline{\xi}_{tk} \sim \mathcal{N}(0, H\nu_k(\delta)\Sigma_{tk}^{-1})$ we obtain w.h.p.:

$$|\phi^\top \overline{\xi}_{tk}| \leqslant \|\phi\|_{\Sigma_{tk}^{-1}} \|\overline{\xi}_{tk}\|_{\Sigma_{tk}} = \sqrt{\gamma_k(\delta)}\|\phi\|_{\Sigma_{tk}^{-1}} \qquad (22)$$

where $\sqrt{\gamma_k(\delta)} = \widetilde{O}((dH)^{3/2})$ is also defined fully in App. D. Note the pseudonoise *worst-case* bound is $\sqrt{Hd}$ worse than the corresponding environment noise.

5.3 Stochastic Optimism and Random Walk

We now want to show that OPT-RLSVI injects enough pseudonoise that the estimated value function $\overline{V}_{1k}(s_{1k})$ at the initial state s_{1k} is optimistic with constant probability (see App. F). We call this event $\mathcal{O}_k$:

$$\mathcal{O}_k \stackrel{def}{=} \left\{ (\overline{V}_{1k} - V_1^\star)(s_{1k}) \geqslant 0 \right\}. \qquad (23)$$

Note that the optimal policy $\pi^\star$ maximizes $Q^\star$ and not the $\overline{Q}$ computed by the algorithm and thus

$$(\overline{V}_{1k} - V_1^\star)(s_{1k}) \geqslant (\overline{Q}_{1k} - Q_1^\star)(s_{1k}, \pi_1^\star(s_{1k})). \quad (24)$$

Now, the goal is to leverage Eqn. (14) to inductively expand this inequality by unrolling a trajectory under the policy $\pi^\star$. To access the result in Eqn. (14) we need to have $\|\phi_1(s_{1k}, \pi_1^\star(s_{1k}))\|_{\Sigma_{1k}^{-1}} \leqslant \alpha_L$. For now, we just assume that this is the case to motivate the idea. In that case, applying Eqn. (14) gives us

$$(\overline{V}_{1k} - V_1^\star)(s_{1k}) \geqslant \phi_1(s_{1k}, \pi_1^\star(s_{1k}))^\top (\overline{\xi}_{1k} + \overline{\eta}_{1k})$$
$$+ \mathbb{E}_{s'|s_{1k}, \pi_1^\star(s_{1k})}[(\overline{V}_{2k} - V_2^\star)(s')]. \qquad (25)$$

Now we can inductively apply the same reasoning to the term inside of the expectation (assuming that we always get features with small Σ^{-1}-norm). Using x_t to denote the states sampled under $\pi^\star$ to avoid confusion with s_{tk} observed by the algorithm, we get

$$\geqslant \sum_{t=1}^{H} \mathbb{E}_{x_t \sim \pi^\star | s_{1k}} \left[\phi_t(x_t, \pi_t^\star(x_t))^\top (\overline{\xi}_{tk} + \overline{\eta}_{tk}) \right] \qquad (26)$$

Since these trajectories over x come from $\pi^\star$ and the environment, they do not depend on the algorithm's policy and with respect to the pseudonoise

$\overline{\xi}$, they are non-random. If we let $\phi_t^\star$ denote $\mathbb{E}_{x_t \sim \pi^\star | s_{1k}} \phi_t(x_t, \pi^\star(x_t))$, and apply Eqn. (21) we get with probability at least $1 - \delta$ that:

$$\sum_{t=1}^{H} (\phi_t^\star)^\top (\overline{\xi}_{tk} + \overline{\eta}_{tk}) \geqslant \sum_{t=1}^{H} [(\phi_t^\star)^\top \overline{\xi}_{tk} - \sqrt{\nu_k(\delta)} \|\phi_t^\star\|_{\Sigma_{tk}^{-1}}]$$

$$\geqslant \sum_{t=1}^{H} (\phi_t^\star)^\top \overline{\xi}_{tk} - \sqrt{H\nu_k(\delta)} \left(\sum_{t=1}^{H} \|\phi_t^\star\|_{\Sigma_{tk}^{-1}}^2 \right)^{1/2} \quad (27)$$

where the second inequality is Cauchy-Schwarz.

Note that the only randomness in this quantity comes from the pseudonoise we inject. We can think of this sum as a one-dimensional normal random walk over H steps with a negative bias. Moreover, if we chose each $\overline{\xi}_{tk} \sim \mathcal{N}(0, H\nu_k(\delta)\Sigma_{tk}^{-1})$, we know that

$$\sum_{t=1}^{H} (\phi_t^\star)^\top \overline{\xi}_{tk} \sim \mathcal{N}\left(0, \sum_{t=1}^{H} H\nu_k(\delta)\|\phi_t^\star\|_{\Sigma_{tk}^{-1}}^2 \right). \quad (28)$$

Comparing this with Eqn. (27) we can immediately see that the standard deviation of the sum of pseudonoise terms is exactly the bound on the bias induced by the high probability bound on the sum of the environment noise $\overline{\eta}_{tk}$. Thus we can conclude that

$$\mathbf{P}\left((\overline{V}_{1k} - V_1^\star)(s_{1k}) \geqslant 0 \right) \geqslant \Phi(-1) \quad (29)$$

where Φ is the normal CDF. This is just the result that we are looking for. However, this presentation avoided the technicalities of handling the cases where $\|\phi_t(x_t, \pi_t^\star(x_t))\|_{\Sigma_{tk}^{-1}} > \alpha_L$ and $\overline{Q}_{tk}$ takes the default value. At a high level the default value is optimistic and so it cannot reduce the probability of optimism. This is handled carefully in Lem. F.1 and F.2 of the appendix, where we obtain a recursion structurally similar to Eqn. (27) albeit with a less interpretable definition of $\phi_t^\star$. One important detail is that our choice of when to default does not depend on the $\overline{\xi}_{tk}$ and is thus non-random with respect to the pseudonoise.

5.4 High Probability Regret Bound

In this section we provide a high level sketch of the main argument that allows us to obtain a high probability regret bound for OPT-RLSVI under Asm. 1. In particular, we assume that the "good event" holds, which lets us use the bounds in Eqn. (21) and (22).

First, we recall the definition of regret up to episode K from the preliminaries and further add and subtract the randomized value functions $\overline{V}_{1k}$ to get that REGRET(K) decomposes as

$$\sum_{k=1}^{K} \left(\underbrace{V_1^\star - \overline{V}_{1k}}_{\text{Pessimism}} + \underbrace{\overline{V}_{1k} - V_1^{\pi_k}}_{\text{Estimation}} \right)(s_{1k}) \quad (30)$$

5.4.1 Bound on estimation

We need to distinguish between cases where $\|\phi_{tk}\|_{\Sigma_{tk}^{-1}} \leqslant \alpha_L$, which we will denote by $\mathcal{S}_{tk}$ for small feature, or not, which we will denote by $\mathcal{S}_{tk}^c$ for its complement. Under $\mathcal{S}_{tk}$ linearity of the representation can be used via Eqn. (14) and under $\mathcal{S}_{tk}^c$ we can use the trivial upper bound of H on the difference in values:

$$\left(\overline{V}_{1k} - V_1^{\pi_k}\right)(s_{1k}) \leqslant H\mathbb{1}\{\mathcal{S}_{1k}^c\} + \quad (31)$$
$$\left(\phi_{1k}^\top \left(\overline{\xi}_{1k} + \overline{\eta}_{1k}\right) + \underbrace{\mathbb{E}_{s'|s_{1k}, a_{1k}}[\left(\overline{V}_{2k} - V_2^{\pi_k}\right)(s')]}_{=\dot{\zeta}_{1k} + (\overline{V}_{2k} - V_2^{\pi_k})(s_{2k})} \right) \mathbb{1}\{\mathcal{S}_{1k}\}$$

where $\dot{\zeta}_{tk} \overset{def}{=} \mathbb{1}\{\mathcal{S}_{1k}\}\left(\mathbb{E}_{s'|s_{tk}, a_{tk}} \left(\overline{V}_{t+1,k} - V_{t+1}^{\pi_k}\right)(s') - \left(\overline{V}_{t+1,k} - V_{t+1}^{\pi_k}\right)(s_{t+1,k}) \right)$ is a bounded martingale difference sequence on the good event. Induction and summing over k eventually yields:

$$\leqslant \sum_{k=1}^{K} \sum_{t=1}^{H} \underbrace{H\mathbb{1}\{\mathcal{S}_{tk}^c\}}_{\text{Warmup}} + \underbrace{\phi_{tk}^\top \left(\overline{\xi}_{tk} + \overline{\eta}_{tk}\right)\mathbb{1}\{\mathcal{S}_{tk}\}}_{\text{Linear Regime}} + \underbrace{\dot{\zeta}_{tk}\mathbb{1}\{\mathcal{S}_{tk}\}}_{\text{Martingale}}.$$

The martingale term can be bounded with high probability by $\tilde{O}(H\sqrt{T})$ using Azuma-Hoeffding.

The first term measures regret during "warmup", when the algorithm cannot guarantee that the value function estimates are bounded and needs to use the default function. In Lem. G.5 we bound it and obtain:

$$\tilde{O}\left(\frac{H^2 d}{\alpha_L^2} \right) = \tilde{O}\left(H^5 d^4 \right) \quad (32)$$

which is $\sqrt{T}$-free and is thus a lower order term.

For the dominant linear regime term we can use the high probability bounds from Eqn. (21) and (22) along with two applications of Cauchy-Schwarz:

$$\leqslant \sum_{k=1}^{K} \sum_{t=1}^{H} \|\phi_{tk}\|_{\Sigma_{tk}^{-1}} \left(\sqrt{\gamma_k(\delta)} + \sqrt{\nu_k(\delta)} \right) \quad (33)$$

$$\leqslant \sqrt{K} \times \sum_{t=1}^{H} \underbrace{\sqrt{\sum_{k=1}^{K} \|\phi_{tk}\|_{\Sigma_{tk}^{-1}}^2}}_{\tilde{O}(\sqrt{d})} \times \left(\underbrace{\sqrt{\gamma_K(\delta)}}_{\tilde{O}(H^{3/2}d^{3/2})} + \underbrace{\sqrt{\nu_K(\delta)}}_{\tilde{O}(Hd)} \right)$$

This final bound on the sum of the squared norm of the features is a standard quantity that arises in linear bandit computations (Abbasi-Yadkori et al., 2011). We can see that the estimation term gives the same regret bound reported in the Thm. 1. Now we show that the pessimism term is of the same order.

5.4.2 Bound on Pessimism

For optimistic algorithms the pessimism term of the regret $\sum_{k=1}^{K}(V_1^\star - \overline{V}_{1k}^{\pi_k})(s_{1k})$ is negative by construction; here we need to work a little more. As seen

above, the algorithm has at least a *constant* probability of being optimistic. When it is, it makes progress similar to a deterministic optimistic algorithm, and when it is not, it is still choosing a reasonable policy (using shrinking confidence intervals) so that the mistakes it makes become less and less severe. Ultimately, we would like to transform the pessimism term into an estimation argument that we can handle as before. So, we first upper bound $V_1^\star$ and then lower bound $\overline{V}_{1k}$ by randomized value functions with specific choices for the pseudonoise. As more samples are collected, the pseudonoise shrinks and the estimates converge.

Upper Bound on $V_1^\star$. Consider drawing $\widetilde{\xi}_{tk}$'s defined as independent and identically distributed copies of the $\overline{\xi}_{tk}$'s. Let $\widetilde{\mathcal{O}}_k$ be the event that in episode k the algorithm obtains an optimistic value function $\widetilde{V}_{1k}$ using these $\widetilde{\xi}_{tk}$ in place of $\overline{\xi}_{tk}$. Explicitly,

$$\widetilde{\mathcal{O}}_k = \{(\widetilde{V}_{1k} - V_{1k}^\star)(s_{1k}) \geqslant 0\}. \tag{34}$$

Note that since the $\widetilde{\xi}_{tk}$ are iid copies of the $\overline{\xi}_{tk}$ we have that $\mathbf{P}(\widetilde{\mathcal{O}}_k)$ is equal to $\mathbf{P}(\mathcal{O}_k) = \Phi(-1)$ from Sec. 5.3. Taking conditional expectation $\mathbb{E}_{\widetilde{\xi}|\widetilde{\mathcal{O}}_k}$ over the $\widetilde{\xi}_{tk}$ for $t \in [H]$ gives us an upper bound:

$$V_{1k}^\star(s_{1k}) \leqslant \mathbb{E}_{\widetilde{\xi}|\widetilde{\mathcal{O}}_k} \widetilde{V}_{1k}(s_{1k}) \tag{35}$$

by definition of the event $\widetilde{\mathcal{O}}_k$.

Lower Bound on $\overline{V}_{1k}$. Under the high probability bound on the pseudonoise of Eqn. (22) we consider the below optimization program over the *optimization variables* ξ_{tk}'s, which are constrained to satisfy the same bound on the pseudonoise of Eqn. (22):

$$\min_{\{\xi_{tk}\}_{t=1,\dots,H}} V_{1k}^\xi(s_{1k}) \tag{36}$$
$$\|\xi_{tk}\|_{\Sigma_{tk}} \leqslant \sqrt{\gamma_k(\delta)}, \quad \forall t \in [H]$$

where V_{1k}^ξ is analogous to $\overline{V}_{1k}$ derived from our algorithm, but with the optimization variables ξ_{tk} in place of $\overline{\xi}_{tk}$. Solving the program above would give a value function $\underline{V}_{1k}$ such that:

$$\underline{V}_{1k}(s_{1k}) \leqslant \overline{V}_{1k}(s_{1k}) \tag{37}$$

whenever the $\overline{\xi}_{tk}$'s obey the high probability bound.

Putting it together. Now we chain the upper bound of Eqn. (35) with the lower bound of Eqn. (37):

$$\left(V_{1k}^\star - \overline{V}_{1k}\right)(s_{1k}) \leqslant \mathbb{E}_{\widetilde{\xi}|\widetilde{\mathcal{O}}_k}[(\widetilde{V}_{1k} - \underline{V}_{1k})(s_{1k})]. \tag{38}$$

We can connect this conditional expectation with the probability of optimism to get to a concentration

bound by applying the law of total expectation:

$$\mathbb{E}_{\widetilde{\xi}}[(\widetilde{V}_{1k} - \underline{V}_{1k})(s_{1k})] = \mathbb{E}_{\widetilde{\xi}|\widetilde{\mathcal{O}}_k}[(\widetilde{V}_{1k} - \underline{V}_{1k})(s_{1k})]\,\mathbf{P}(\widetilde{\mathcal{O}}_k)$$
$$+ \underbrace{\mathbb{E}_{\widetilde{\xi}|\widetilde{\mathcal{O}}_k^c}[(\widetilde{V}_{1k} - \underline{V}_{1k})(s_{1k})]\,\mathbf{P}(\widetilde{\mathcal{O}}_k^c)}_{\geqslant 0}. \tag{39}$$

This inequality holds by the same reasoning as Eqn. (37) with high probability since the $\widetilde{\xi}_{tk}$ are also in the set over which $\underline{V}_{1k}$ is minimized. Dividing by $\mathbf{P}(\widetilde{\mathcal{O}}_k)$ and chaining with Eqn. (38) gives us:

$$\left(V_{1k}^\star - \overline{V}_{1k}\right)(s_{1k}) \leqslant \mathbb{E}_{\widetilde{\xi}}[(\widetilde{V}_{1k} - \underline{V}_{1k})(s_{1k})]/\mathbf{P}(\widetilde{\mathcal{O}}_k).$$

Now, since the $\widetilde{\xi}_{tk}$ are iid copies of the $\overline{\xi}_{tk}$ that the algorithm computes we have that $\mathbb{E}_{\widetilde{\xi}}[\widetilde{V}_{1k}(s_{1k})] = \mathbb{E}_{\overline{\xi}}[\overline{V}_{1k}(s_{1k})]$ and $\mathbf{P}(\mathcal{O}_k) = \mathbf{P}(\widetilde{\mathcal{O}}_k)$. So we can define a martingale difference sequence $\ddot{\zeta}_k \overset{def}{=} \mathbb{E}_{\widetilde{\xi}}[\widetilde{V}_{1k}(s_{1k})] - \overline{V}_{1k}(s_{1k})$ and get our final bound on the pessimism as:

$$\left(V_{1k}^\star - \overline{V}_{1k}\right)(s_{1k}) \leqslant \frac{(\overline{V}_{1k} - \underline{V}_{1k})(s_{1k}) + \ddot{\zeta}_k}{\mathbf{P}(\mathcal{O}_k)}. \tag{40}$$

When summing over the episodes $k \in [K]$, the martingale can be bounded with high probability by Azuma-Hoeffding as $\sum_{k=1}^K \ddot{\zeta}_k = \widetilde{O}(H\sqrt{K})$. To bound the remaining term we add and subtract $V_1^{\pi_k}$ to get:

$$\left(\sum_{k=1}^K [(\overline{V}_{1k} - V_1^{\pi_k})(s_{1k}) + (V_1^{\pi_k} - \underline{V}_{1k})(s_{1k})]\right)/\mathbf{P}(\mathcal{O}_k).$$

Each of these is bounded by arguments similar to those in Sec. 5.4.1. We discuss this in detail in Lem. G.4.

It is instructive to re-examine Eqn. (40), ignoring the martingale term. While the left hand side is negative for optimistic algorithms, for OPT-RLSVI it is upper bounded by a difference in estimated value functions (which shrinks with more data) times the inverse probability of being optimistic $1/\mathbf{P}(\mathcal{O}_k)$. In other words, roughly once every $1/\mathbf{P}(\mathcal{O}_k)$ episodes the algorithm is optimistic and exploration progress is made.

6 Concluding Remarks

This work proposes the first high probability regret bounds for (a modified version of) RLSVI with function approximation, confirming its sound exploration principles. Perhaps unsurprisingly, we inherit an extra $\sqrt{dH}$ regret factor compared to an optimistic approach which can be explained by analogy to the bandit literature. Whether Thompson sampling-based algorithms need to suffer this extra factor compared to their optimistic counterparts remains a fundamental research question in exploration. Our work enriches the literature on provably efficient exploration algorithms with function approximation with a new algorithmic design as well as a new set of analytical techniques.

References

Yasin Abbasi-Yadkori, David Pal, and Csaba Szepesvari. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

Marc Abeille, Alessandro Lazaric, et al. Linear thompson sampling revisited. *Electronic Journal of Statistics*, 11(2):5165–5197, 2017.

Shipra Agrawal and Navin Goyal. Further optimal regret bounds for thompson sampling. In *AISTATS*, volume 31 of *JMLR Workshop and Conference Proceedings*, pages 99–107. JMLR.org, 2013.

Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NIPS)*, pages 1184–1194. Curran Associates, Inc., 2017.

Mohammad Gheshlaghi Azar, Ian Osband, and Remi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2017.

Ronen I. Brafman and Moshe Tennenholtz. R-MAX - A general polynomial time algorithm for near-optimal reinforcement learning. *J. Mach. Learn. Res.*, 3:213–231, 2002.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *NIPS*, pages 2249–2257, 2011.

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

Christoph Dann, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. On oracle-efficient pac rl with rich observations. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1429–1439, 2018.

Christoph Dann, Lihong Li, Wei Wei, and Emma Brunskill. Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning*, pages 1507–1516, 2019.

Simon S Du, Sham M Kakade, Ruosong Wang, and Lin F Yang. Is a good representation sufficient for sample efficient reinforcement learning? *arXiv preprint arXiv:1910.03016*, 2019.

Amir-massoud Farahmand, Csaba Szepesvári, and Rémi Munos. Error propagation for approximate policy and value iteration. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.

Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Ronald Ortner. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. 2018.

Gene H Golub and Charles F Van Loan. *Matrix Computations*. JHU Press, 2012.

Aditya Gopalan and Shie Mannor. Thompson sampling for learning parameterized markov decision processes. In *COLT*, volume 40 of *JMLR Workshop and Conference Proceedings*, pages 861–898. JMLR.org, 2015.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 2010.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E. Schapire. Contextual decision processes with low Bellman rank are PAC-learnable. In Doina Precup and Yee Whye Teh, editors, *International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 1704–1713, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/jiang17c.html.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. *arXiv preprint arXiv:1907.05388*, 2019.

K. Lakshmanan, Ronald Ortner, and Daniil Ryabko. Improved regret bounds for undiscounted continuous reinforcement learning. In *ICML*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 524–532. JMLR.org, 2015.

Tor Lattimore and Csaba Szepesvari. Learning with good feature representations in bandits and in rl with a generative model. *arXiv preprint arXiv:1911.07676*, 2019.

Rémi Munos. Error bounds for approximate value iteration. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2005.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(May):815–857, 2008.

Ronald Ortner and Daniil Ryabko. Online regret bounds for undiscounted continuous reinforcement learning. In *NIPS*, pages 1772–1780, 2012.

Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In *International Conference on Machine Learning (ICML)*, 2017.

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped DQN. In *Advances in Neural Information Processing Systems (NIPS)*, 2016a.

Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. In *International Conference on Machine Learning (ICML)*, 2016b.

Yi Ouyang, Mukul Gagrani, Ashutosh Nayyar, and Rahul Jain. Learning unknown markov decision processes: A thompson sampling approach. In *NIPS*, pages 1333–1342, 2017.

David Pollard. Empirical processes: theory and applications. In *NSF-CBMS regional conference series in probability and statistics*, pages i–86. JSTOR, 1990.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1994. ISBN 0471619779.

Daniel Russo. Worst-case regret bounds for exploration via randomized value functions. *arXiv preprint arXiv:1906.02870*, 2019.

Steven E Shreve and Dimitri P Bertsekas. Alternative theoretical frameworks for finite horizon discrete-time stochastic optimal control. *SIAM Journal on control and optimization*, 16(6):953–978, 1978.

Benjamin Van Roy and Shi Dong. Comments on the du-kakade-wang-yang lower bounds. *arXiv preprint arXiv:1911.07910*, 2019.

Lin F Yang and Mengdi Wang. Reinforcement leaning in feature space: Matrix bandit, kernels, and regret bound. *arXiv preprint arXiv:1905.10389*, 2019a.

Lin F Yang and Mengdi Wang. Sample-optimal parametric q-learning with linear transition models. *arXiv preprint arXiv:1902.04779*, 2019b.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning (ICML)*, 2019. URL `http://proceedings.mlr.press/v97/zanette19a.html`.

Appendix

Table of Contents

A   Notation

We provide this table for easy reference. Notation will also be defined as it is introduced.

We denote with H the episode length, with K the total number of episodes, and with $T = HK$ the time elapsed. We denote with $k \in [K]$ the current episode, with $t \in [H]$ the current timestep. We use the subscript tk to indicate the quantity at timestep t of episode k and $t + 1, k$ for the subsequent step.

Table 1: Symbols

s_{tk}	$\stackrel{def}{=}$	state encountered in timestep t of episode k
a_{tk}	$\stackrel{def}{=}$	action taken by the algorithm in timestep t of episode k
ϕ_{tk}	$\stackrel{def}{=}$	$\phi_t(s_{tk}, a_{tk})$
r_{tk}	$\stackrel{def}{=}$	$r_t(s_{tk}, a_{tk})$
λ	$\stackrel{def}{=}$	regularization parameter
Σ_{tk}	$\stackrel{def}{=}$	$\sum_{i=1}^{k-1} \phi_{ti}\phi_{ti}^{\top} + \lambda I$
$\widehat{\theta}_{tk}$	$\stackrel{def}{=}$	$\Sigma_{tk}^{-1} \left(\sum_{i=1}^{k-1} \phi_{ti}[r_{ti} + \overline{V}_{t+1,k}(s_{t+1,i})] \right)$
$\sqrt{\beta_k(\delta)}$	$\stackrel{def}{=}$	$c_1 H d \sqrt{\log \left(\frac{Hdk \max(1, L_\phi)\max(1, L_\psi)\max(1, L_r)\lambda}{\delta} \right)}$
$\sqrt{\nu_k(\delta)}$	$\stackrel{def}{=}$	$\sqrt{\beta_k(\delta)} + \sqrt{\lambda}L_\phi(3HL_\psi + L_r) + 4\epsilon H\sqrt{dk}$
$\sqrt{\gamma_k(\delta)}$	$\stackrel{def}{=}$	$c_2\sqrt{dH\nu_k(\delta)\log(d/\delta)}$
$\overline{\xi}_{tk}$	$\stackrel{def}{=}$	Pseudonoise distributed as $\mathcal{N}(0, H\nu_k(\delta)\Sigma_{tk}^{-1})$
$\overline{\theta}_{tk}$	$\stackrel{def}{=}$	$\widehat{\theta}_{tk} + \overline{\xi}_{tk}$
α_U	$\stackrel{def}{=}$	$\frac{1}{4(\sqrt{\gamma_k(\delta)})}$
α_L	$\stackrel{def}{=}$	$\frac{\alpha_U}{2}$
$\overline{Q}_{tk}(s, a)$	$\stackrel{def}{=}$	$\begin{cases} \phi^{\top}\overline{\theta}_{tk}, & \text{if } \|\phi_t(s,a)\|_{\Sigma_{tk}^{-1}} \leqslant \alpha_L \\ H - t + 1, & \text{if } \|\phi_t(s,a)\|_{\Sigma_{tk}^{-1}} \geqslant \alpha_U \\ \frac{\alpha_U - \|\phi_t(s,a)\|_{\Sigma_{tk}^{-1}}}{\alpha_U - \alpha_L}\left(\phi^{\top}\overline{\theta}_{tk}\right) + \frac{\|\phi_t(s,a)\|_{\Sigma_{tk}^{-1}} - \alpha_L}{\alpha_U - \alpha_L}(H - t + 1), & \text{otherwise} \end{cases}$
$\overline{V}_{tk}(s)$	$\stackrel{def}{=}$	$\max_a \overline{Q}_{tk}(s, a)$
$\pi_k(s)$	$\stackrel{def}{=}$	policy executed by the algorithm in episode k, i.e. $\arg\max_a \overline{Q}_{tk}(s, a)$
$\mathcal{S}_{tk}$	$\stackrel{def}{=}$	Event $\left\{ \|\phi_{tk}\|_{\Sigma_{tk}^{-1}} \leqslant \alpha_L \right\}$
$\mathcal{S}_{tk}^c$	$\stackrel{def}{=}$	Event $\left\{ \|\phi_{tk}\|_{\Sigma_{tk}^{-1}} > \alpha_L \right\}$ (complement of $\mathcal{S}_{tk}$)
L_ϕ	$\stackrel{def}{=}$	upper bound on $\|\phi\|$
L_ψ	$\stackrel{def}{=}$	upper bound on $\int_s \|\psi_t(s')\|$ for all $t \in [H]$
L_r	$\stackrel{def}{=}$	upper bound on $\|\theta_r\|$
L_θ	$\stackrel{def}{=}$	upper bound on $\|\theta_t^\pi\|$ (equal to $L_r + (H - 1)L_\psi$)
$\Delta_t^P(\cdot|s, a)$	$\stackrel{def}{=}$	$\mathbb{P}_t(\cdot|s, a) - \phi(s, a)_t^{\top}\psi_t(\cdot)$
$\Delta_t^r(s, a)$	$\stackrel{def}{=}$	$r_t(s, a) - \phi(s, a)_t^{\top}\theta_t^r$
ϵ	$\stackrel{def}{=}$	bound on $|\Delta_t^r(s, a)|$ and $\|\Delta_t^P(\cdot|s, a)\|_1$
$\overline{\eta}_{tk}$	$\stackrel{def}{=}$	$\Sigma_{tk}^{-1}\sum_{i=1}^{k-1}\phi_{ti}\left(\overline{V}_{t+1,k}(s_{t+1,i}) - \mathbb{E}_{s'|s_{it}, a_{it}}[\overline{V}_{t+1,k}(s')]\right)$

$$\overline{\lambda}_{tk}^{\pi} \quad \overset{def}{=} \quad -\lambda \Sigma_{tk}^{-1} \left(\int_{s'} \psi_t(s')(\overline{V}_{t+1,k} - V_{t+1}^{\pi})(s') + \theta_t^{\pi} \right)$$

$$\Delta_t^{\pi}(s,a) \quad \overset{def}{=} \quad Q_t^{\pi}(s,a) - \phi_t(s,a)^{\top}\theta_t^{\pi}$$

$$\overline{m}_{tk}^{\pi} \quad \overset{def}{=} \quad \phi_t(s,a)^{\top}\Sigma_{tk}^{-1}\sum_{i=1}^{k-1}\phi_{ti}\left[\Delta_t^r(s_{ti},a_{ti}) + \int_{s'}\Delta_t^P(s'|s_{ti},a_{ti})\overline{V}_{t+1,k}(s')\right] + \Delta_t^{\pi}(s,a)$$

$$\qquad\qquad\qquad - \int_{s'}\Delta_t^P(s'|s,a)(\overline{V}_{t+1,k} - V_{t+1}^{\pi})(s')$$

$$\mathcal{H}_{tk} \quad \overset{def}{=} \quad \{s_{ij}, a_{ij}, r_{ij} : j \leqslant k, \quad i \leqslant t \text{ if } j = k \text{ else } i \leqslant H\}$$

$$\overline{\mathcal{H}}_{tk} \quad \overset{def}{=} \quad \mathcal{H}_{Hk} \bigcup \{\overline{\xi}_{ik} : i \geqslant t\}$$

$$\mathcal{G}_{tk}^{\overline{\xi}} \quad \overset{def}{=} \quad \left\{ |\phi_t(s,a)^{\top}\overline{\xi}_{tk}| \leqslant \sqrt{\gamma_k(\delta)}\|\phi_t(s,a)\|_{\Sigma_{tk}^{-1}} \right\}$$

$$\mathcal{G}_{tk}^{\overline{\eta}} \quad \overset{def}{=} \quad \left\{ |\phi_t(s,a)^{\top}\overline{\eta}_{tk}| \leqslant \sqrt{\beta_k(\delta)}\|\phi_t(s,a)\|_{\Sigma_{tk}^{-1}} \right\}$$

$$\mathcal{G}_{tk}^{\overline{\lambda}} \quad \overset{def}{=} \quad \left\{ \forall \pi, \quad |\phi_t(s,a)^{\top}\overline{\lambda}_{tk}^{\pi}| \leqslant \sqrt{\lambda}L_{\phi}(3HL_{\psi} + L_r)\|\phi_t(s,a)\|_{\Sigma_{tk}^{-1}} \right\}$$

$$\mathcal{G}_{tk}^{\overline{m}} \quad \overset{def}{=} \quad \left\{ \forall \pi, \quad |\overline{m}_{tk}^{\pi}(s,a)| \leqslant 4\epsilon H(\sqrt{dk}\|\phi_t(s,a)\|_{\Sigma_{tk}^{-1}} + 1) \right\}$$

$$\mathcal{G}_{tk}^{\overline{Q}} \quad \overset{def}{=} \quad \left\{ \forall s,a, \quad |(\overline{Q}_{tk} - Q_t^{\star})(s,a)| \leqslant H - t + 1 \right\}$$

$$\overline{\mathcal{G}}_{tk} \quad \overset{def}{=} \quad \{\mathcal{G}_{tk}^{\overline{\xi}} \cap \mathcal{G}_{tk}^{\overline{\eta}} \cap \mathcal{G}_{tk}^{\overline{\lambda}} \cap \mathcal{G}_{tk}^{\overline{m}} \cap \mathcal{G}_{tk}^{\overline{Q}}\}$$

$$\overline{\mathcal{G}}_k \quad \overset{def}{=} \quad \bigcap_{t\in[H]}\overline{\mathcal{G}}_{tk}$$

$$\widetilde{\xi}_{tk} \quad \overset{def}{=} \quad \text{i.i.d. copy of the pseudonoise } \widetilde{\xi}_{tk}, \text{ useful for the regret proof. All overline quantities can}$$

$$\qquad\qquad\qquad \text{be translated to tilde by exchanging pseudonoise variables in the value iteration.}$$

$$\widetilde{\mathcal{O}}_k \quad \overset{def}{=} \quad \left\{ \left(\widetilde{V}_{1k} - V_1^{\star}\right)(s_{1k}) \geqslant -4H^2\epsilon \right\}$$

B Assumptions

In this section we formally present the main assumption that the MDP is approximately low-rank and show that the definition immediately implies the existence of approximately linear Q functions for any policy. Moreover, the corresponding parameters to these Q functions have bounded norm.

Assumption B.1 (ϵ-approximate low-rank MDP). *(Jin et al., 2019; Yang and Wang, 2019a) For any $\epsilon \leqslant 1$, an MDP $(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$ is ϵ-approximate low-rank with feature maps $\phi_t : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ if for every $t \in [H]$ there exists an unknown function $\psi_t : \mathcal{S} \to \mathbb{R}^d$ and an unknown vector $\theta_t^r \in \mathbb{R}^d$ such that*

$$\|\mathbb{P}_t(\cdot|s,a) - \phi(s,a)_t^\top \psi_t(\cdot)\|_1 \leqslant \epsilon, \qquad |r_t(s,a) - \phi(s,a)_t^\top \theta_t^r| \leqslant \epsilon. \tag{41}$$

Moreover assume the bounds

1. *$\|\phi_t(s,a)\| \leqslant L_\phi$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $t \in [H]$.*

2. *$\int_{\mathcal{S}} \|\psi_t(s)\| \leqslant L_\psi$ for all $t \in [H]$.*

3. *$\|\theta_t^r\| \leqslant L_r$ for all $t \in [H]$.*

Definition B.2 (Misspecification). *We can define the following misspecification quantities*

$$\Delta_t^P(\cdot|s,a) \stackrel{def}{=} \mathbb{P}_t(\cdot|s,a) - \phi(s,a)_t^\top \psi_t(\cdot), \qquad \|\Delta_t^P(\cdot|s,a)\|_1 = \int_{s'} |\Delta_t^P(s'|s,a)| \leqslant \epsilon \tag{42}$$

$$\Delta_t^r(s,a) \stackrel{def}{=} r_t(s,a) - \phi(s,a)_t^\top \theta_t^r, \qquad |\Delta_t^r(s,a)| \leqslant \epsilon \tag{43}$$

where the inequalities follow from the Assumption B.1.

Corollary B.3 (Linear Q functions). *For any policy π, there exist some $\theta_t^\pi \in \mathbb{R}^d$ for all $t \in [H]$ such that for all s, a*

$$|Q_t^\pi(s,a) - \phi(s,a)_t^\top \theta_t^\pi| \leqslant (H - t + 1)\epsilon. \tag{44}$$

Moreover, $\|\theta_t^\pi\| \leqslant L_r + (H - t)L_\psi \stackrel{def}{=} L_\theta$.

Proof. Since $Q_t^\pi(s,a) = \phi(s,a)^\top \left(\theta_t^r + \int \psi(s') V_{t+1}^\pi(s') ds'\right)$, we set

$$\theta_t^\pi = \theta_t^r + \int_{s'} \psi_t(s') V_{t+1}^\pi(s') \tag{45}$$

Note that by the assumption that the rewards are in $[0, 1]$ the true value functions V_t^π are always in $[0, H - t + 1]$. By the triangle inequality and Bellman equation followed by an application of Definition B.2

$$|Q_t^\pi(s,a) - \phi_t(s,a)^\top \theta_t^\pi| \leqslant |r_t(s,a) - \phi(s,a)_t^\top \theta_t^r| + \left|\mathbb{E}_{s'|s,a}[V_{t+1}^\pi(s')] - \phi_t(s,a)^\top \int_{s'} \psi_t(s') V_{t+1}^\pi(s')\right| \tag{46}$$

$$\leqslant \epsilon + \left|\int_{s'} (P_t(s'|s,a) - \phi_t(s,a)^\top \psi_t(s')) V_{t+1}^\pi(s')\right| \tag{47}$$

$$\leqslant \epsilon + \|V_{t+1}^\pi\|_\infty \|\Delta_t^P(\cdot|s,a)\|_1 \leqslant \epsilon + (H - t)\epsilon = (H - t + 1)\epsilon \tag{48}$$

To prove the second part of the statement, note that by the triangle inequality and Assumption B.1

$$\|\theta_t^\pi\| \leqslant \|\theta_t^r\| + \|\int_{s'} \psi_t(s') V_{t+1}^\pi(s')\| \leqslant L_r + \|V_{t+1}^\pi\|_\infty L_\psi \leqslant L_r + (H - t)L_\psi. \tag{49}$$

$\square$

Definition B.4 (Optimal parameters). *We can denote the parameters associated with the optimal policy $\pi^\star$ as $\theta_t^\star = \theta_t^{\star,P} + \theta_t^r$.*

C Decomposition of Unclipped Q-values

In this section we prove the main decomposition lemma that will be useful throughout. The lemma decomposes the difference between the function defined by the estimated $\overline{\theta}_{tk}$ and the true Q^π for any policy π into several parts: the expected difference of corresponding value functions at the next state, the projected environment noise, the pseudonoise, a term due to the regularizer λ and a term due to the misspecification (i.e. the ϵ error) of the low-rank MDP.

These terms are defined in the following notation:

$$\overline{\eta}_{tk} \stackrel{def}{=} \Sigma_{tk}^{-1} \sum_{i=1}^{k-1} \phi_{ti}\left(\overline{V}_{t+1,k}(s_{t+1,i}) - \mathbb{E}_{s'|s_{ti},a_{it}}[\overline{V}_{t+1,k}(s')]\right) \tag{50}$$

$$\overline{\lambda}_{tk}^\pi \stackrel{def}{=} -\lambda\Sigma_{tk}^{-1}\left(\int_{s'} \psi_t(s')(\overline{V}_{t+1,k} - V_{t+1}^\pi)(s') + \theta_t^\pi\right) \tag{51}$$

$$\Delta_t^\pi(s,a) \stackrel{def}{=} Q_t^\pi(s,a) - \phi_t(s,a)^\top\theta_t^\pi \tag{52}$$

$$\overline{m}_{tk}^\pi(s,a) \stackrel{def}{=} \phi_t(s,a)^\top\Sigma_{tk}^{-1}\sum_{i=1}^{k-1}\phi_{ti}\left[\Delta_t^r(s_{ti},a_{ti}) + \int_{s'}\Delta_t^P(s'|s_{ti},a_{ti})\overline{V}_{t+1,k}(s')\right] + \Delta_t^\pi(s,a) \tag{53}$$

$$- \int_{s'}\Delta_t^P(s'|s,a)(\overline{V}_{t+1,k} - V_{t+1}^\pi)(s') \tag{54}$$

Lemma C.1 (Decomposition of unclipped Q-values). *For $t \in [H]$ and any policy π:*

$$\phi_t(s,a)^\top\overline{\theta}_{tk} - Q_t^\pi(s,a) = \mathbb{E}_{s'|s,a}\left[\left(\overline{V}_{t+1,k} - V_{t+1}^\pi\right)(s')\right] + \phi_t(s,a)^\top(\overline{\eta}_{tk} + \overline{\xi}_{tk} + \overline{\lambda}_{tk}^\pi) + \overline{m}_{tk}^\pi(s,a) \tag{55}$$

where $\mathbb{E}_{s'|s,a}[\cdot] = \mathbb{E}_{s'\sim\mathbb{P}_t(\cdot|s,a)}[\cdot]$ and the index t will be clear from context.

Proof. By Corollary B.3 we have:

$$\phi_t(s,a)^\top\overline{\theta}_{tk} - Q_t^\pi(s,a) = \phi_t(s,a)^\top(\overline{\theta}_{tk} - \theta_t^\pi) + \Delta_t^\pi(s,a) \tag{56}$$

By substituting the definition of $\overline{\theta}_{tk}$ and the linear regression, we get:

$$= \phi_t(s,a)^\top\left(\overline{\xi}_{tk} + \underbrace{\Sigma_{tk}^{-1}\sum_{i=1}^{k-1}\phi_{ti}(r_{ti} + \overline{V}_{t+1,k}(s_{t+1,i}))}_{=\hat{\theta}_{tk}} - \theta_t^\pi\right) + \Delta_t^\pi(s,a) \tag{57}$$

Moving θ_t^π inside the sum by multiplying by $\Sigma_{tk}^{-1}\Sigma_{tk} = I$ we get

$$= \phi(s,a)^\top\left(\overline{\xi}_{tk} + \Sigma_{tk}^{-1}\left(-\lambda\theta_t^\pi + \sum_{i=1}^{k-1}\phi_{ti}\left(r_{ti} + \overline{V}_{t+1,k}(s_{t+1,i}) - \phi_{ti}^\top\theta_t^\pi\right)\right)\right) + \Delta_t^\pi(s,a). \tag{58}$$

Now we expand $\phi_{ti}^\top\theta_t^\pi = \phi_{ti}^\top(\theta_t^r + \int_{s'}\psi(s')V_{t+1}^\pi(s'))$ (see Eq. 45)

$$= \phi_t(s,a)^\top\left(\overline{\xi}_{tk} + \Sigma_{tk}^{-1}\left(-\lambda\theta_t^\pi + \sum_{i=1}^{k-1}\phi_{ti}\left[r_{ti} + \overline{V}_{t+1,k}(s_{t+1,i}) - \phi_{ti}^\top\left(\theta_t^r + \int_{s'}\psi_t(s')V_{t+1}^\pi(s')\right)\right]\right)\right) + \Delta_t^\pi(s,a). \tag{59}$$

Next we add and subtract $\mathbb{E}_{s'|s_{ti},a_{ti}}[\overline{V}_{t+1,k}(s')] - \phi_{ti}^\top \int_{s'} \psi(s')\overline{V}_{t+1,k}(s')$ and rearrange terms to get

$$= \phi_t(s,a)^\top \left(\overline{\xi}_{tk} - \lambda\Sigma_{tk}^{-1}\theta_t^\pi + \Sigma_{tk}^{-1} \sum_{i=1}^{k-1} \phi_{ti} \underbrace{\left[r_{ti} - \phi_{ti}^\top\theta_t^r + \mathbb{E}_{s'|s_{ti},a_{ti}}\left[\overline{V}_{t+1,k}(s')\right] - \phi_{ti}^\top \int_{s'} \psi_t(s')\overline{V}_{t+1,k}(s') \right]}_{=\Delta_t^r(s_{ti},a_{ti}) + \int_{s'} \Delta_t^P(s'|s_{ti},a_{ti})\overline{V}_{t+1,k}(s')} \right. \tag{60}$$

$$+ \Sigma_{tk}^{-1} \sum_{i=1}^{k-1} \phi_{ti} \underbrace{\left[\overline{V}_{t+1,k}(s_{t+1,i}) - \mathbb{E}_{s'|s_{ti},a_{ti}}\left[\overline{V}_{t+1,k}(s')\right] \right]}_{\overline{\eta}_{tk}} \tag{61}$$

$$+ \left. \Sigma_{tk}^{-1} \sum_{i=1}^{k-1} \phi_{ti} \left[\phi_{ti}^\top \int_{s'} \psi_t(s')(\overline{V}_{t+1,k} - V_{t+1}^\pi)(s') \right] \right) + \Delta_t^\pi(s,a). \tag{62}$$

We can add and subtract a regularizer term and cancel $\Sigma_{tk}^{-1}\Sigma_{tk}$ to get

$$= \phi_t(s,a)^\top \left(\overline{\xi}_{tk} + \overline{\eta}_{tk} + \int_{s'} \psi_t(s')(\overline{V}_{t+1,k} - V_{t+1}^\pi)(s') \right. \tag{63}$$

$$- \lambda\Sigma_{tk}^{-1} \underbrace{\left[\theta_t^\pi + \int_{s'} \psi_t(s')(\overline{V}_{t+1,k} - V_{t+1}^\pi)(s') \right]}_{\overline{\lambda}_{tk}^\pi} \tag{64}$$

$$+ \left. \Sigma_{tk}^{-1} \sum_{i=1}^{k-1} \phi_{ti} \left[\Delta_t^r(s_{ti},a_{ti}) + \int_{s'} \Delta_t^P(s'|s_{ti},a_{ti})\overline{V}_{t+1,k}(s') \right] \right) + \Delta_t^\pi(s,a) \tag{65}$$

Finally we replace the integral by the true expectation plus a misspecification term

$$= \phi_t(s,a)^\top(\overline{\xi}_{tk} + \overline{\eta}_{tk} + \overline{\lambda}_{tk}^\pi) + \mathbb{E}_{s'|s,a}[(\overline{V}_{t+1,k} - V_{t+1}^\pi)(s')] \tag{66}$$

$$\underbrace{- \int_{s'} \Delta_t^P(s'|s,a)(\overline{V}_{t+1,k} - V_{t+1}^\pi)(s') + \phi_t(s,a)^\top\Sigma_{tk}^{-1} \sum_{i=1}^{k-1} \phi_{ti} \left[\Delta_t^r(s_{ti},a_{ti}) + \int_{s'} \Delta_t^P(s'|s_{ti},a_{ti})\overline{V}_{t+1,k}(s') \right] + \Delta_t^\pi(s,a)}_{\overline{m}_{tk}^\pi(s,a)}$$

$$\tag{67}$$

$\square$

D   Defining the Good Event

In this section we formally define the filtrations that compose the history of the algorithm at any point during its runtime. Then we define the values $\beta_k(\delta), \nu_k(\delta)$, and $\gamma_k(\delta)$ that are used to define our high confidence bounds. We use these to choose settings of the cutoff parameters α_L, α_U. Finally, we define the good events whereby the terms from the decomposition presented in the preceding section are bounded in terms of the design matrix and $\beta_k(\delta), \nu_k(\delta)$, and $\gamma_k(\delta)$.

Definition D.1 (Filtrations). *For any $t \in [H]$ and any k define the filtrations*

$$\mathcal{H}_{tk} \stackrel{def}{=} \{s_{ij}, a_{ij}, r_{ij} : j \leqslant k, \quad i \leqslant t \text{ if } j = k \text{ else } i \leqslant H\} \tag{68}$$

$$\mathcal{H}_k \stackrel{def}{=} \mathcal{H}_{H,k} \tag{69}$$

$$\overline{\mathcal{H}}_{tk} \stackrel{def}{=} \mathcal{H}_k \bigcup \{\overline{\xi}_{ik} : i \geqslant t\} \tag{70}$$

$$\overline{\mathcal{H}}_k \stackrel{def}{=} \overline{\mathcal{H}}_{1k} \tag{71}$$

Definition D.2 (Noise bounds). *For any $\delta > 0$ and some constants c_1, c_2 let*

$$\sqrt{\beta_k(\delta)} \stackrel{def}{=} c_1 H d \sqrt{\log\left(\frac{H dk \max(1, L_\phi) \max(1, L_\psi) \max(1, L_r)\lambda}{\delta}\right)} \tag{72}$$

$$\sqrt{\nu_k(\delta)} \stackrel{def}{=} \sqrt{\beta_k(\delta)} + \sqrt{\lambda} L_\phi (3HL_\psi + L_r) + 4\epsilon H \sqrt{dk} \tag{73}$$

$$\sqrt{\gamma_k(\delta)} \stackrel{def}{=} c_2 \sqrt{dH \nu_k(\delta) \log(d/\delta)} \tag{74}$$

Note that this functions are monotonically increasing in k, e.g., $\sqrt{\beta_k(\delta)} \leqslant \sqrt{\beta_{k+1}(\delta)}$.

Definition D.3 (Default cutoff). *Set*

$$\alpha_U \stackrel{def}{=} \frac{1}{4(\sqrt{\gamma_k(\delta)})} \leqslant \frac{1}{2(\sqrt{\nu_k(\delta)} + \sqrt{\gamma_k(\delta)})} \tag{75}$$

$$\alpha_L \stackrel{def}{=} \alpha_U/2 \tag{76}$$

Definition D.4 (Good event). *Define*

$$\mathcal{G}_{tk}^{\overline{\xi}} \stackrel{def}{=} \left\{ |\phi_t(s,a)^\top \overline{\xi}_{tk}| \leqslant \sqrt{\gamma_k(\delta)} \|\phi_t(s,a)\|_{\Sigma_{tk}^{-1}} \right\} \tag{77}$$

$$\mathcal{G}_{tk}^{\overline{\eta}} \stackrel{def}{=} \left\{ |\phi_t(s,a)^\top \overline{\eta}_{tk}| \leqslant \sqrt{\beta_k(\delta)} \|\phi_t(s,a)\|_{\Sigma_{tk}^{-1}} \right\} \tag{78}$$

$$\mathcal{G}_{tk}^{\overline{\lambda}} \stackrel{def}{=} \left\{ \forall \, \pi, \quad |\phi_t(s,a)^\top \overline{\lambda}_{tk}^\pi| \leqslant \sqrt{\lambda} L_\phi (3HL_\psi + L_r) \|\phi_t(s,a)\|_{\Sigma_{tk}^{-1}} \right\} \tag{79}$$

$$\mathcal{G}_{tk}^{\overline{m}} \stackrel{def}{=} \left\{ \forall \, \pi, \quad |\overline{m}_{tk}^\pi(s,a)| \leqslant 4\epsilon H (\sqrt{dk} \|\phi_t(s,a)\|_{\Sigma_{tk}^{-1}} + 1) \right\} \tag{80}$$

$$\mathcal{G}_{tk}^{\overline{Q}} \stackrel{def}{=} \left\{ \forall \, s,a, \quad |(\overline{Q}_{tk} - Q_t^\star)(s,a)| \leqslant H - t + 1 \right\} \tag{81}$$

And then the good events are the intersections

$$\overline{\mathcal{G}}_{tk} \stackrel{def}{=} \{\mathcal{G}_{tk}^{\overline{\xi}} \cap \mathcal{G}_{tk}^{\overline{\eta}} \cap \mathcal{G}_{tk}^{\overline{\lambda}} \cap \mathcal{G}_{tk}^{\overline{m}} \cap \mathcal{G}_{tk}^{\overline{Q}}\} \tag{82}$$

$$\overline{\mathcal{G}}_k \stackrel{def}{=} \bigcap_{t \in [H]} \overline{\mathcal{G}}_{tk} \tag{83}$$

E   Concentration

This section will prove that the good events happen with high probability. The tricky part is showing that the estimates $\overline{Q}_{tk}$ remain nicely bounded. To do this we bound each of the four separate terms (misspecification, regularization, pseudonoise, and environment noise) with high probability when conditioned on bounded $\overline{Q}$ values at time $t+1$. Then we use an inductive argument to show that this means that all terms and the $\overline{Q}$ values are bounded across all timesteps with high probability.

E.1   Bounding the Misspecification Error

Lemma E.1 (Misspecification). *For any t, k, s, a and any policy π, if*

$$\left|(\overline{Q}_{t+1,k} - Q^\star_{t+1})(s,a)\right| \leqslant H - t \tag{84}$$

then

$$|\overline{m}^\pi_{tk}(s,a)| \leqslant 4\epsilon H \left(\sqrt{dk}\|\phi_t(s,a)\|_{\Sigma^{-1}_{tk}} + 1\right) \tag{85}$$

Proof. Recall the definition of $\overline{m}^\pi_{tk}$ in Eq. 53

$$|\overline{m}^\pi_{tk}(s,a)| = \left|\phi_t(s,a)^\top \Sigma^{-1}_{tk} \sum_{i=1}^{k-1} \phi_{ti}\left[\Delta^r_t(s_{ti}, a_{ti}) + \int_{s'} \Delta^P_t(s'|s_{ti}, a_{ti})\overline{V}_{t+1,k}(s')\right] + \Delta^\pi_t(s,a) \right. \tag{86}$$

$$\left. - \int_{s'} \Delta^P_t(s'|s,a)(\overline{V}_{t+1,k} - V^\pi_{t+1})(s')\right|. \tag{87}$$

Under event $\mathcal{G}^{\overline{Q}}_{t+1,k}$, we have that $|(\overline{V}_{t+1,k} - V^\pi_{t+1})(s')| \leqslant |(\overline{V}_{t+1,k} - V^\star_{t+1})(s')| + |(V^\star_{t+1} - V^\pi_{t+1})(s')| \leqslant 2H$. Then, applying the triangle inequality, Holder, and bounds from Definition B.2 and Corollary B.3 as well as previous bound on the estimated value functions, we can erite

$$|\overline{m}^\pi_{tk}(s,a)| \leqslant (\epsilon + \epsilon H)\left|\phi_t(s,a)^\top \Sigma^{-1}_{tk} \sum_{i=1}^{k-1} \phi_{ti}\right| + \epsilon H + 3\epsilon H. \tag{88}$$

Finally, grouping terms and applying Cauchy-Schwarz twice we get

$$\leqslant 4\epsilon H \left(\|\phi_t(s,a)\|_{\Sigma^{-1}_{tk}} \left\|\sum_{i=1}^{k-1} \phi_{ti}\right\|_{\Sigma^{-1}_{tk}} + 1\right) \tag{89}$$

$$\leqslant 4\epsilon H \left(\sqrt{k}\|\phi_t(s,a)\|_{\Sigma^{-1}_{tk}} \left(\sum_{i=1}^{k-1} \|\phi_{ti}\|^2_{\Sigma^{-1}_{tk}}\right)^{1/2} + 1\right). \tag{90}$$

The result follows by applying Lemma I.3. □

E.2   Bounding the Regularization

Lemma E.2 (Regularization). *For any t, k, π and any features $\phi_t(s,a)$, if*

$$\left|(\overline{Q}_{t+1,k} - Q^\star_{t+1})(s,a)\right| \leqslant H - t \tag{91}$$

then

$$|\phi_t(s,a)^\top \overline{\lambda}^\pi_{tk}| \leqslant \sqrt{\lambda}L_\phi(3HL_\psi + L_r)\|\phi_t(s,a)\|_{\Sigma^{-1}_{tk}} \tag{92}$$

Proof. By Cauchy-Schwarz and the fact that the maximal eigenvalue of Σ_{tk}^{-1} is at most $1/\lambda$

$$|\phi_t(s,a)^\top \overline{\lambda}_{tk}^\pi| = \left|\phi_t(s,a)^\top \lambda \Sigma_{tk}^{-1}\left(\int_{s'} \psi_t(s')(\overline{V}_{t+1,k} - V_{t+1}^\pi)(s') + \theta_t^\pi\right)\right| \tag{93}$$

$$\leqslant \sqrt{\lambda}\|\phi_t(s,a)\|_{\Sigma_{tk}^{-1}}\left(\left\|\int_{s'} \psi_t(s')(\overline{V}_{t+1,k} - V_{t+1}^\pi)(s')\right\| + \|\theta_t^\pi\|\right) \tag{94}$$

$$\leqslant \sqrt{\lambda}\|\phi_t(s,a)\|_{\Sigma_{tk}^{-1}}\left(\int_{s'} \left\|\psi_t(s')(\overline{V}_{t+1,k} - V_{t+1}^\pi)(s')\right\| + \|\theta_t^\pi\|\right) \tag{95}$$

$$\leqslant \sqrt{\lambda}\|\phi_t(s,a)\|_{\Sigma_{tk}^{-1}}\left(\int_{s'} \|\psi_t(s')\||(\overline{V}_{t+1,k} - V_{t+1}^\pi)(s')| + \|\theta_t^\pi\|\right) \tag{96}$$

$$\leqslant \sqrt{\lambda}\|\phi_t(s,a)\|_{\Sigma_{tk}^{-1}}\left(\|\overline{V}_{t+1,k} - V_{t+1}^\pi\|_\infty \int_{s'} \|\psi_t(s')\| + \|\theta_t^\pi\|\right) \tag{97}$$

Applying the hypothesis of the lemma and the bounds from Assumption B.1 and Corollary B.3

$$\leqslant \sqrt{\lambda}L_\phi[2HL_\psi + (L_r + (H-t)L_\psi)]\|\phi_t(s,a)\|_{\Sigma_{tk}^{-1}} \leqslant \sqrt{\lambda}L_\phi(3HL_\psi + L_r)\|\phi_t(s,a)\|_{\Sigma_{tk}^{-1}}. \tag{98}$$

$\square$

E.3 Bounding the Environment Noise

Lemma E.3 (Concentration inductive step). *Fix t and k. For any $\delta > 0$ and conditioned for all s,a and all $z > t$ on*

$$\left|(\overline{Q}_{z,k} - Q_z^\star)(s,a)\right| \leqslant H - t \tag{99}$$

and on

$$\|\overline{\xi}_{t+1,k}\|_{\Sigma_{t+1,k}} \leqslant \sqrt{\gamma_k(\delta)} \tag{100}$$

then with probability at least $1 - \delta$

$$|\phi_t(s,a)^\top \overline{\eta}_{tk}| \leqslant \sqrt{\beta_k(\delta)}\|\phi_t(s,a)\|_{\Sigma_{tk}^{-1}} \tag{101}$$

Proof. Recall the definition of $\overline{\eta}_{tk}$ given in Eq. 50. By Cauchy-Schwarz:

$$|\phi_t(s,a)^\top \overline{\eta}_{tk}| \leqslant \|\phi_t(s,a)\|_{\Sigma_{tk}^{-1}}\|\overline{\eta}_{tk}\|_{\Sigma_{tk}^{-1}} \tag{102}$$

where

$$\|\overline{\eta}_{tk}\|_{\Sigma_{tk}^{-1}} = \left\|\sum_{i=1}^{k-1} \phi_{ti}\left(\overline{V}_{t+1,k}(s_{t+1,i}) - \mathbb{E}_{s'|s_{ti},a_{ti}}[\overline{V}_{t+1,k}(s')]\right)\right\|_{\Sigma_{tk}^{-1}} \tag{103}$$

First, we will show that given the hypothesis of the lemma, we can bound

$$\|\overline{\theta}_{t+1,k}\| \leqslant 2H\sqrt{kd/\lambda} + \sqrt{\gamma_k(\delta)/\lambda} \tag{104}$$

To see this, note that $\|\overline{V}_{t+2,k}\|_\infty \leqslant 2(H-t-1)$ from Eq. 99 and so applying Cauchy-Schwarz gives us

$$\|\widehat{\theta}_{t+1,k}\| = \|\Sigma_{t+1,k}^{-1}\sum_{i=1}^{k-1}\phi_{ti}(r_{t+1,i} + \overline{V}_{t+2,k}(s_{t+2,i}))\| \leqslant \|\Sigma_{t+1,k}^{-1/2}\|\|\sum_{i=1}^{k-1}\phi_{t+1,i}(r_{t+1,i} + \overline{V}_{t+2,k}(s_{t+2,i}))\|_{\Sigma_{t+1,k}^{-1}} \tag{105}$$

$$\leqslant \frac{1}{\sqrt{\lambda}}\sqrt{k}\left(\sum_{i=1}^{k-1}\|\phi_{t+1,i}(r_{t+1,i} + \overline{V}_{t+2,k}(s_{t+2,i}))\|_{\Sigma_{t+1,k}^{-1}}^2\right)^{1/2} \tag{106}$$

$$\leqslant \frac{1}{\sqrt{\lambda}}(2(H-t-1)+1)\sqrt{k}\left(\sum_{i=1}^{k-1}\|\phi_{ti}\|_{\Sigma_{tk}^{-1}}^2\right)^{1/2} \tag{107}$$

$$\leqslant 2H\sqrt{kd/\lambda} \tag{108}$$

where the last inequality comes from Lemma I.3. With this bound in hand, we can now proceed with a covering argument over the functions $\overline{V}_{t+1,k}$ to bound $\overline{\eta}_{tk}$.

For any $\theta \in \mathbb{R}^d$ with $\|\theta\| \leqslant 2H\sqrt{kd/\lambda} + \sqrt{\gamma_k(\delta)/\lambda}$ and $\Sigma \in \mathbb{R}^{d \times d}$ symmetric and positive definite with $\|\Sigma\| \leqslant \frac{1}{\lambda}$, we define

$$Q_t^{\theta,\Sigma}(s,a) \stackrel{def}{=} \begin{cases} \phi_t(s,a)^\top \theta, & \text{if } \|\phi_t(s,a)\|_\Sigma \leqslant \alpha_L \\ H - t + 1 & \text{if } \|\phi_t(s,a)\|_\Sigma \geqslant \alpha_U \\ \left(\frac{\alpha_U - \|\phi_t(s,a)\|_\Sigma}{\alpha_U - \alpha_L}\right)\phi_t(s,a)^\top \theta + \left(\frac{\|\phi_t(s,a)\|_\Sigma - \alpha_L}{\alpha_U - \alpha_L}\right)(H - t + 1) & \text{otherwise} \end{cases} \tag{109}$$

Let $V^{\theta,\Sigma}$ be the corresponding value function. Note that $\overline{V}_{t+1,k} = V^{\overline{\theta}_{t+1,k}, \Sigma_{t+1,k}^{-1}}$.

Define

$$O_{t+1} \stackrel{def}{=} \left\{ \theta, \Sigma : \|\theta\| \leqslant 2H\sqrt{kd/\lambda} + \sqrt{\gamma_k(\delta)/\lambda}, \quad \|\Sigma\| \leqslant \frac{1}{\lambda}, \quad |(Q_{t+1}^{\theta,\Sigma} - Q_{t+1}^\star)(s,a)| \leqslant H - t \ \ \forall \ s, a \right\} \tag{110}$$

So that by the hypothesis of the lemma, $\overline{\theta}_{t+1,k}, \Sigma_{t+1,k}^{-1} \in O_{t+1}$.

For any $(\theta, \Sigma) \in O_{t+1}$ and $i \in [k-1]$ define

$$x_i^{\theta,\Sigma} \stackrel{def}{=} V^{\theta,\Sigma}(s_{t+1,i}) - \mathbb{E}_{s'|s_{ti},a_{ti}}[V^{\theta,\Sigma}(s')] \tag{111}$$

Then x_i defines a martingale difference sequence with filtration $\mathcal{H}_{ti}$. Moreover, by the definition of O_{t+1}, each x_i is bounded in absolute value by $2H$ (from last condition in (110)) so that each x_i is a $2H$-subgaussian random variable.

So, by Lemma I.1 the $x_i^{\theta,\Sigma}$ induce a self normalizing process so that

$$\left\| \sum_{i=1}^{k-1} \phi_i x_i^{\theta,\Sigma} \right\|_{\Sigma_{tk}^{-1}} \leqslant 4H\left(d\log\left(\frac{kL_\phi^2 + \lambda}{\lambda}\right) + \log(1/\delta) \right)^{1/2} \tag{112}$$

Note that the ε-covering number of O_{t+1} as a Euclidean ball in $\mathbb{R}^{d+d^2}$ of radius $2H\sqrt{kd/\lambda} + \sqrt{\gamma_k(\delta)/\lambda} + 1/\lambda$, denoted $N_\varepsilon(O_{t+1})$, is bounded by Lemma I.5 as $(3(2H\sqrt{kd/\lambda} + \sqrt{\gamma_k(\delta)/\lambda} + 1/\lambda)/\varepsilon)^{d^2+d}$. So, by a union bound, with probability at least $1 - \delta$ we have for all $(\theta, \Sigma) \in O_{t+1}$ that

$$\left\| \sum_{i=1}^{k-1} \phi_i x_i^{\theta,\Sigma} \right\|_{\Sigma_{tk}^{-1}} \leqslant 4H\left(d\log\left(\frac{kL_\phi^2 + \lambda}{\lambda}\right) + \log(N_\varepsilon(O_{t+1})/\delta) \right)^{1/2} \tag{113}$$

$$\leqslant 4H\left(d\log\left(\frac{kL_\phi^2 + \lambda}{\lambda}\right) + (d^2 + d)\log\left(3(2H\sqrt{kd/\lambda} + \sqrt{\gamma_k(\delta)/\lambda} + 1/\lambda)/\varepsilon\right) + \log(1/\delta) \right)^{1/2} \tag{114}$$

$$\leqslant 8Hd\left(\log\left(\frac{kL_\phi^2 + \lambda}{\lambda}\right) + \log\left(3(2H\sqrt{kd/\lambda} + \sqrt{\gamma_k(\delta)/\lambda} + 1/\lambda)/\varepsilon\right) + \log(1/\delta) \right)^{1/2} \tag{115}$$

To conclude the proof, we choose a specific $(\theta, \Sigma) \in O_{t+1}$ such that $\|\theta - \overline{\theta}_{t+1,k}\| \leqslant \varepsilon$ and $\|\Sigma - \Sigma_{t+1,k}^{-1}\|_F \leqslant \varepsilon$. Then

$$\|\eta_{tk}\|_{\Sigma_{tk}^{-1}} = \left\| \sum_{i=1}^{k-1} \phi_i x_i^{\overline{\theta}_{t+1,k}, \Sigma_{t+1,k}^{-1}} \right\|_{\Sigma_{tk}^{-1}} \tag{116}$$

$$\leqslant \left\| \sum_{i=1}^{k-1} \phi_i x_i^{\theta,\Sigma} \right\|_{\Sigma_{tk}^{-1}} + \left\| \sum_{i=1}^{k-1} \phi_i (x_i^{\theta,\Sigma} - x_i^{\overline{\theta}_{t+1,k}, \Sigma_{t+1,k}^{-1}}) \right\|_{\Sigma_{tk}^{-1}} \tag{117}$$

Then we can bound

$$\left\| \sum_{i=1}^{k-1} \phi_i (x_i^{\theta,\Sigma} - x_i^{\overline{\theta}_{t+1,k}, \Sigma_{t+1,k}^{-1}}) \right\|_{\Sigma_{tk}^{-1}} \leqslant kL_\phi \sup_i \left| x_i^{\theta,\Sigma} - x_i^{\overline{\theta}_{t+1,k}, \Sigma_{t+1,k}^{-1}} \right| \tag{118}$$

Plugging in the definition of the x_i and applying Lemma E.4 we bound

$$\sup_i \left| x_i^{\theta,\Sigma} - x_i^{\overline{\theta}_{t+1,k},\Sigma_{t+1,k}^{-1}} \right| = \sup_i \left| (V^{\theta,\Sigma} - \overline{V}_{t+1,i})(s_{t+1,i}) - \mathbb{E}_{s'|s_{ti},a_{ti}}[(V^{\theta,\Sigma} - \overline{V}_{t+1,i})(s')] \right| \tag{119}$$

$$\leqslant 2 \sup_{s,a} |(Q^{\theta,\Sigma} - \overline{Q}_{t+1,k})(s,a)| \tag{120}$$

$$\leqslant 2\sqrt{\varepsilon} \frac{L_\phi(4H^2)}{\alpha_U - \alpha_L} \tag{121}$$

So we can bound the covering error by 1 if we choose ε small enough such that

$$\varepsilon \leqslant \left(\frac{\alpha_U - \alpha_L}{8kL_\phi^2 H^2} \right)^2 \tag{122}$$

Then with probability at least $1 - \delta$, combining (115) with (117), I.6, and the choice of ε we get

$$\|\overline{\eta}_{tk}\|_{\Sigma_{tk}^{-1}} \leqslant \left\| \sum_{i=1}^{k-1} \phi_i x_i^{\theta,\Sigma} \right\|_{\Sigma_{tk}^{-1}} + 1 \leqslant \sqrt{\beta_k(\delta)} \tag{123}$$

as desired. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad\square$

Lemma E.4 (Covering Lemma). *This lemma uses the notation defined within the previous lemma, suppressing indices. Take (θ,Σ) and (θ',Σ') in O (see Eq. 110 for generic t) such that $\|\theta - \theta'\| \leqslant \varepsilon$ and $\|\Sigma - \Sigma'\| \leqslant \varepsilon$ with $\varepsilon \leqslant \min\{1, \frac{H}{3L_\phi}, \frac{\alpha_U - \alpha_L}{L_\phi^2}\}$, then*

$$\sup_{s,a} |(Q^{\theta,\Sigma} - Q^{\theta',\Sigma'})(s,a)| \leqslant \sqrt{\varepsilon} \frac{L_\phi(4H^2)}{\alpha_U - \alpha_L} \tag{124}$$

Proof. Note that by the assumption, for any ϕ with $\|\phi\| \leqslant L_\phi$

$$\left| \|\phi\|_\Sigma - \|\phi\|_{\Sigma'} \right| = \left| \sqrt{\phi^\top \Sigma \phi} - \sqrt{\phi^\top \Sigma' \phi} \right| \leqslant \sqrt{|\phi^\top(\Sigma - \Sigma')\phi|} \leqslant \sqrt{\|\phi\|\|(\Sigma - \Sigma')\|\|\phi\|} \leqslant \sqrt{\varepsilon}L_\phi \tag{125}$$

Now we need to split into cases. Since θ, Σ and θ', Σ' are interchangable, the following 5 cases cover all possibilities.

Case 1 (linear-linear): $\|\phi(s,a)\|_\Sigma \leqslant \alpha_L$ and $\|\phi(s,a)\|_{\Sigma'} \leqslant \alpha_L$.

We can apply Cauchy-Schwarz and the definition of the case to get

$$|(Q^{\theta,\Sigma} - Q^{\theta',\Sigma'})(s,a)| = |\phi(s,a)^\top(\theta - \theta')| \leqslant L_\phi \varepsilon \tag{126}$$

Case 2 (linear-interpolating): $\|\phi(s,a)\|_\Sigma \leqslant \alpha_L$ and $\alpha_L \leqslant \|\phi(s,a)\|_{\Sigma'} \leqslant \alpha_L + \sqrt{\varepsilon}L_\phi \leqslant \alpha_U$

Applying (125) and the definition of the case,

$$\|\phi(s,a)\|_{\Sigma'} \leqslant \|\phi(s,a)\|_\Sigma + |\|\phi(s,a)\|_{\Sigma'} - \|\phi(s,a)\|_\Sigma| \leqslant \alpha_L + \sqrt{\varepsilon}L_\phi. \tag{127}$$

Moreover, by our choice of $\theta, \Sigma \in O$ which induces bounded Q functions we can bound

$$|\phi(s,a)^\top \theta - (H-t)| \leqslant |\phi(s,a)^\top \theta| + H \leqslant 3H \tag{128}$$

So if we set

$$q' \stackrel{def}{=} \frac{\|\phi(s,a)\|_{\Sigma'} - \alpha_L}{\alpha_U - \alpha_L} \leqslant \frac{\alpha_L + \sqrt{\varepsilon}L_\phi - \alpha_L}{\alpha_U - \alpha_L} = \frac{\sqrt{\varepsilon}L_\phi}{\alpha_U - \alpha_L} \tag{129}$$

then we have by the triangle inequality, equation (126), and the above reasoning,

$$|(Q^{\theta,\Sigma} - Q^{\theta',\Sigma'})(s,a)| = |\phi(s,a)^\top \theta - (1-q')\phi(s,a)^\top \theta' - q'(H-t)| \tag{130}$$

$$\leqslant (1-q')|\phi(s,a)^\top(\theta - \theta')| + q'|\phi(s,a)^\top \theta - (H-t)| \tag{131}$$

$$\leqslant (1-q')L_\phi \varepsilon + q'|\phi(s,a)^\top \theta - (H-t)| \tag{132}$$

$$\leqslant L_\phi \varepsilon + \frac{\sqrt{\varepsilon}L_\phi(3H)}{\alpha_U - \alpha_L} \tag{133}$$

Case 3 (default-default): $\alpha_U \leqslant \|\phi(s,a)\|_\Sigma$ and $\alpha_U \leqslant \|\phi(s,a)\|_{\Sigma'}$.

Then we have that

$$|(Q^{\theta,\Sigma} - Q^{\theta',\Sigma'})(s,a)| = |(H-t) - (H-t)| = 0. \tag{134}$$

Case 4 (default-interpolating): $\alpha_U \leqslant \|\phi(s,a)\|_\Sigma$ and $\alpha_L \leqslant \alpha_U - \sqrt{\varepsilon}L_\phi \leqslant \|\phi(s,a)\|_{\Sigma'} \leqslant \alpha_U$

By the definition of the case

$$-\sqrt{\varepsilon}L_\phi \leqslant \|\phi(s,a)\|_{\Sigma'} - \alpha_U, \tag{135}$$

so that defining q' as before

$$1 - q' = 1 - \frac{\|\phi(s,a)\|_{\Sigma'} - \alpha_L}{\alpha_U - \alpha_L} = \frac{\alpha_U - \|\phi(s,a)\|_{\Sigma'}}{\alpha_U - \alpha_L} \leqslant \frac{\sqrt{\varepsilon}L_\phi}{\alpha_U - \alpha_L}. \tag{136}$$

And thus, applying (136) and (128) again we get

$$|(Q^{\theta,\Sigma} - Q^{\theta',\Sigma'})(s,a)| = |(H-t) - (1-q')\phi(s,a)^\top \theta' - q'(H-t)| \tag{137}$$

$$\leqslant (1-q')|\phi(s,a)^\top \theta' - (H-t)| \tag{138}$$

$$\leqslant \frac{\sqrt{\varepsilon}L_\phi(3H)}{\alpha_U - \alpha_L} \tag{139}$$

Case 5 (interpolating-interpolating): $\alpha_L \leqslant \|\phi(s,a)\|_{\Sigma'} \leqslant \alpha_U$ and $\alpha_L \leqslant \|\phi(s,a)\|_{\Sigma'} \leqslant \alpha_U$

Letting q be analogous to q' but for Σ and applying (125) we have

$$|q - q'| = \frac{|\|\phi(s,a)\|_\Sigma - \alpha_L - (\|\phi(s,a)\|_{\Sigma'} - \alpha_L)|}{\alpha_U - \alpha_L} \leqslant \frac{\sqrt{\varepsilon}L_\phi}{\alpha_U - \alpha_L} \tag{140}$$

Thus we have that

$$|(Q^{\theta,\Sigma} - Q^{\theta',\Sigma'})(s,a)| = |(1-q)\phi(s,a)^\top \theta + q(H-t) \tag{141}$$

$$- (1-q')\phi(s,a)^\top \theta' - q'(H-t)| \tag{142}$$

$$\leqslant \frac{\sqrt{\varepsilon}L_\phi(3H)}{\alpha_U - \alpha_L} \left(|\phi(s,a)^\top(\theta - \theta')| + H\right) \tag{143}$$

$$\leqslant \frac{\sqrt{\varepsilon}L_\phi(3H)}{\alpha_U - \alpha_L} \left(L_\phi \varepsilon + H\right) \leqslant \frac{\sqrt{\varepsilon}L_\phi(4H^2)}{\alpha_U - \alpha_L} \tag{144}$$

Taking the max over all of the cases (which is case 5) yields the result. $\qquad\square$

E.4 Bounding the Q values

Lemma E.5 (Boundedness inductive step). *Assume that $\epsilon < \frac{1}{10H}$ and that for all s,a*

$$\left|(\overline{Q}_{t+1,k} - Q_{t+1}^\star)(s,a)\right| \leqslant H - t \tag{145}$$

$$\left|\phi_t(s,a)^\top(\overline{\eta}_{tk} + \overline{\xi}_{tk} + \overline{\lambda}_{tk}^\star) + \overline{m}_{tk}^\star(s,a)\right| \leqslant (\sqrt{\nu_k} + \sqrt{\gamma_k})\|\phi_t(s,a)\|_{\Sigma_{tk}^{-1}} + 4\epsilon H \tag{146}$$

where $\overline{\lambda}_{tk}^{\star}$ *and* $\overline{m}_{tk}^{\star}$ *are as in* (51) *and* (53) *with* $\pi = \pi^{\star}$, *then for all* s, a

$$\left|(\overline{Q}_{tk} - Q_t^{\star})(s, a)\right| \leqslant H - t + 1 \tag{147}$$

Proof. There are two cases, depending on whether the features are large.

Case 1 (large features): $\|\phi_t(s, a)\|_{\Sigma_{tk}^{-1}} \geqslant \alpha_U$.

Then by the definition of $\overline{Q}_{tk}$ from the algorithm (see (109) or Definition 1), we have $0 \leqslant \overline{Q}_{tk}(s, a) \leqslant H - t + 1$. Since $Q_t^{\star}$ must be in the same range, we immediately get

$$\left|(\overline{Q}_{tk} - Q_t^{\star})(s, a)\right| \leqslant H - t + 1 \tag{148}$$

Case 2 (small features): $\|\phi_t(s, a)\|_{\Sigma_{tk}^{-1}} \leqslant \alpha_L$.

In this case we get $\overline{Q}_{tk}(s, a) = \phi_t(s, a)^{\top} \overline{\theta}_{tk}$. So we apply Lemma C.1 to get

$$\left|(\overline{Q}_{tk} - Q_t^{\star})(s, a)\right| = \left| \mathbb{E}_{s'|s,a}\left[\left(\overline{V}_{t+1,k} - V_{t+1}^{\star}\right)(s')\right] + \phi_t(s, a)^{\top}(\overline{\eta}_{tk} + \overline{\xi}_{tk} + \overline{\lambda}_{tk}^{\star}) + \overline{m}_{tk}^{\star}(s, a)\right|. \tag{149}$$

We can split the terms by the triangle inequality. Using the inductive hypothesis (146) gives us

$$\leqslant H - t + (\sqrt{\nu_k} + \sqrt{\gamma_k}) \underbrace{\|\phi_t(s, a)\|_{\Sigma_{tk}^{-1}}}_{\leqslant \alpha_L \text{ since Case 2 holds}} + 4\epsilon H. \tag{150}$$

Finally, by our choice of α_L (see Definition D.3) and using $\epsilon < \frac{1}{10H}$ we get the final bound

$$\leqslant H - t + 1. \tag{151}$$

Case 3 (medium features): $\alpha_L \leqslant \|\phi_t(s, a)\|_{\Sigma_{tk}^{-1}} \leqslant \alpha_U$.

This case immediately follows from applying the first two cases and our choice of α_U (see Definition D.3) along with noting that for any Q^1, Q^2

$$|qQ^1(s, a) + (1 - q)Q^2(s, a) - Q_t^{\star}(s, a)| \leqslant q|(Q^1 - Q_t^{\star})(s, a)| + (1 - q)|(Q^2 - Q_t^{\star})(s, a)| \tag{152}$$

So that when both Q^1, Q^2 satisfy the desired relationship to $Q^{\star}$, so does their interpolation. $\square$

E.5 Putting it all together: good event with high probability

Lemma E.6 (Good event probability). *With* $\epsilon < \frac{1}{10H}$, *for any* K *and any* $\delta > 0$, *with probability* $1 - \delta$ *we have* $\bigcap_{k \leqslant K} \overline{\mathcal{G}}_k$.

Proof. For each k we will induct backwards over t using the preceding lemmas to prove that $\overline{\mathcal{G}}_{tk}$ occurs for all $t \in [H]$ with probability at least $1 - \delta'$. In the following, let $\delta' = \delta/K$ and $\delta'' = \delta'/2H$.

As the base case, consider step H. Since we define $\overline{Q}_{H+1,k} = 0 = Q_{H+1}^{\star}$, we can invoke Lemmas E.1 and E.2 to get $\mathcal{G}_{Hk}^{\overline{\lambda}}$ and $\mathcal{G}_{Hk}^{\overline{m}}$. Then we can apply Lemma I.4 so that and $\mathcal{G}_{Hk}^{\overline{\xi}}$ occurs with probability $1 - \delta''$. Then we can invoke Lemma E.3 to get that conditioned on all these other events we get $\mathcal{G}_{Hk}^{\overline{\eta}}$ with probability at least $1 - \delta''$. Thus, we get the intersection of these events $\{\mathcal{G}_{Hk}^{\overline{\xi}} \cap \mathcal{G}_{Hk}^{\overline{\eta}} \cap \mathcal{G}_{Hk}^{\overline{\lambda}} \cap \mathcal{G}_{Hk}^{\overline{m}}\}$ with probability at least $(1 - \delta'')^2$. Finally, conditioned on $\{\mathcal{G}_{H+1,k}^{\overline{Q}} \cap \mathcal{G}_{Hk}^{\overline{\xi}} \cap \mathcal{G}_{Hk}^{\overline{\eta}} \cap \mathcal{G}_{Hk}^{\overline{\lambda}} \cap \mathcal{G}_{Hk}^{\overline{m}}\}$ we can invoke Lemma E.5 (using the condition on ϵ) to get $\mathcal{G}_{Hk}^{\overline{Q}}$. Combining, we see that $P(\overline{\mathcal{G}}_{Hk}) \geqslant (1 - \delta'')^2$. The inductive step follows the same outline so that conditioning on $\overline{\mathcal{G}}_{tk}$ we have $P(\overline{\mathcal{G}}_{t-1,k}|\overline{\mathcal{G}}_{tk}) \geqslant (1 - \delta'')^2$. Thus, we can bound

$$P\left(\overline{\mathcal{G}}_k\right) \geqslant (1 - \delta'')^{2H} \geqslant 1 - \delta' \tag{153}$$

A union bound over $k \in [K]$ gives the result. $\square$

F   Optimism

In this section we discuss how Algorithm 1 can ensure optimism, and we use $\widetilde{\xi}$ instead of $\overline{\xi}$ to indicate the pseudonoise. This facilitates the proof of Lemma G.4 later, but *the reader should think of the $\widetilde{\xi}$'s as independent and identically distributed copies of the $\overline{\xi}$'s* (therefore with the same 'properties').

To discuss optimism, in Lemma F.1 we lower bound the value function difference by a one-dimensional random walk. The idea is to look at the probability that the algorithm is optimistic along the optimal policy $\pi^\star$. If condition $\|\phi_t(x_t, \pi_t^\star(x_t))\|_{\Sigma_{tk}^{-1}} \leqslant \alpha_L$ was true at every x_t encountered upon following the optimal policy $\pi^\star$, the random variable in the random walk that we obtain would be the projection of the pseudonoise $\overline{\xi}$ along the average feature ϕ encountered upon following $\pi^\star$. In fact, since $\|\phi_t(x_t, \pi_t^\star(x_t))\|_{\Sigma_{tk}^{-1}} \leqslant \alpha_L$ does not always hold, we end up not projecting on the average ϕ_t but on a different $\phi_t^\star$; importantly, this $\phi_t^\star$ is a *non-random* quantity when conditioned on the history $\mathcal{H}_k$ and starting state s_{1k}. This allows us to show optimism by looking at properties of a normal random walk in lemma F.2.

Lemma F.1 (Optimistic Recursion). *Condition on the starting state s_{1k}, the history $\widetilde{\mathcal{H}}_k$ (which is $\overline{\mathcal{H}}_k$ with $\widetilde{\xi}_{tk}$ in place of $\overline{\xi}_{tk}$), and the good event $\widetilde{\mathcal{G}}_k$ (again with $\widetilde{\xi}_{tk}$ in place of $\overline{\xi}_{tk}$). Then for every timestep $t \in [H]$ there exists vector $\phi_t^\star \in \mathbb{R}^d$ that does not depend on any $\widetilde{\xi}_{tk}$ such that:*

$$\left(\widetilde{V}_1 - V_1^\star\right)(s_{1k}) \geqslant \sum_{t=1}^{H} [(\phi_t^\star)^\top \widetilde{\xi}_{tk} - \sqrt{\nu_k(\delta)}\|\phi_t^\star\|_{\Sigma_{tk}^{-1}}] - 4H^2\epsilon. \tag{154}$$

Proof. The proof proceed by induction, and is split into sections.

We will use x rather than s to emphasize the difference between states sampled with $\pi^\star$ (denoted by x) from those sampled with our policy π_k (denoted by s). Before to proceed, recall the definition of $\widetilde{Q}$ (same as $\overline{Q}$) from (109) or (12).

Definitions.   Recursively define the following functions $w_t : \mathcal{S} \to \mathbb{R}$ and $\mathring{w}_t : \mathcal{S} \to \mathbb{R}$, which will be used to define $\phi_t^\star$:

$$w_{t+1}(x_{t+1}) = \int_{\mathcal{S}} \mathring{w}_t(x_t)\mathbb{P}_t(x_{t+1}|x_t, \pi_t^\star(x_t))dx_t \tag{155}$$

$$\mathring{w}_t(x_t) = \begin{cases} w_t(x_t), & \text{if } \|\phi_t(s_t, \pi_t^\star(s_t))\|_{\Sigma_{tk}^{-1}} \leqslant \alpha_L \\ \frac{\alpha_U - \|\phi_t(x_t, \pi_t^\star(x_t))\|_{\Sigma_{tk}^{-1}}}{\alpha_U - \alpha_L} w_t(x_t), & \text{if } \alpha_L < \|\phi_t(x_t, \pi_t^\star(x_t))\|_{\Sigma_{tk}^{-1}} < \alpha_U \\ 0, & \text{if } \|\phi_t(x_t, \pi_t^\star(x_t))\|_{\Sigma_{tk}^{-1}} \geqslant \alpha_U \end{cases} \tag{156}$$

$$w_1(x_1) = 1 \tag{157}$$

$$x_1 = s_{1k} \tag{158}$$

Then we can define

$$\phi_t^\star \stackrel{def}{=} \int_{\mathcal{S}} w_t(x_t)\phi_t(x_t, \pi_t^\star(x_t))dx_t. \tag{159}$$

Importantly, this choice of $\phi_t^\star$ has no dependence on any $\widetilde{\xi}_{tk}$ with $t \in [H]$.

First we prove by induction that the w_t's are positive and integrate to less than 1 for all $t \in [H]$:

$$w_t(x_t) \geqslant 0, \ \forall x_t \in \mathcal{S}$$

$$\int_{S} w_t(x_t)dx_t \leqslant 1 \tag{160}$$

Positivity is immediate from the definition of equation (158) since all quantities are positive. For the integral, assume by induction that at step t it holds that $\int_S w_t(x_t)dx_t \leqslant 1$. For $t+1$ we have:

$$\int_{\mathcal{S}} w_{t+1}(x_{t+1})dx_{t+1} = \int_{\mathcal{S}} \left(\int_{\mathcal{S}} \mathring{w}_t(x_t)\mathbb{P}_t(x_{t+1}|x_t, \pi_t^\star(x_t))dx_t \right) dx_{t+1} \tag{161}$$

$$\stackrel{F}{=} \int_{\mathcal{S}} \left(\int_{\mathcal{S}} \mathring{w}_t(x_t)\mathbb{P}_t(x_{t+1}|x_t, \pi_t^\star(x_t))dx_{t+1} \right) dx_t \tag{162}$$

$$= \int_{\mathcal{S}} \mathring{w}_t(x_t) \underbrace{\left(\int_{\mathcal{S}} \mathbb{P}_t(x_{t+1}|x_t, \pi_t^\star(x_t))dx_{t+1} \right)}_{=1} dx_t \tag{163}$$

$$\leqslant \int_{\mathcal{S}} w_t(x_t)dx_t \leqslant 1. \tag{164}$$

In the last equality we used that $\mathring{w}_t \leqslant w_t$ pointwise (this follows directly by the definition), while step F is due to Fubini's theorem for changing the order of integration.

Starting the main recursion. Let $\mathcal{L}_t$, $\mathcal{M}_t$ $\mathcal{S}_t$ be the event that the norm of the feature evaluated at x_t and the optimal policy is large and small, respectively (x_t is the random variable):

$$\mathcal{S}_t \stackrel{def}{=} \left\{ x_t : \|\phi_t(x_t, \pi_t^\star(x_t))\|_{\Sigma_{tk}^{-1}} \leqslant \alpha_L \right\} \tag{165}$$

$$\mathcal{M}_t \stackrel{def}{=} \left\{ x_t : \alpha_L < \|\phi_t(x_t, \pi_t^\star(x_t))\|_{\Sigma_{tk}^{-1}} < \alpha_U \right\} \tag{166}$$

$$\mathcal{L}_t \stackrel{def}{=} \left\{ x_t : \|\phi_t(x_t, \pi_t^\star(x_t))\|_{\Sigma_{tk}^{-1}} \geqslant \alpha_U \right\} \tag{167}$$

First consider integrating over the state space with respect to $w_t(\cdot)$ the value function difference over the trajectories at step t (the lower bound below holds for every term inside the expectation because $\pi^\star$ is the optimal policy on $Q^\star$ but not necessarily on $\widetilde{Q}$):

$$\int_{\mathcal{S}} w_t(x_t) \left(\widetilde{V}_t - V_t^\star \right) (x_t)dx_t \geqslant \int_{\mathcal{S}} w_t(x_t) \left(\widetilde{Q}_t(x_t, \pi_t^\star(x_t)) - Q_t^\star(x_t, \pi_t^\star(x_t)) \right) dx_t \tag{168}$$

and then partition the statespace $\mathcal{S}$:

$$= \underbrace{\int_{\mathcal{S}_t} w_t(x_t) \left(\widetilde{Q}_t - Q_t^\star \right) (x_t, \pi_t^\star(x_t))dx_t}_{S} + \underbrace{\int_{\mathcal{M}_t} w_t(x_t) \left(\widetilde{Q}_t - Q_t^\star \right) (x_t, \pi_t^\star(x_t))dx_t}_{M} \tag{169}$$

$$+ \underbrace{\int_{\mathcal{L}_t} w_t(x_t) \left(\widetilde{Q}_t - Q_t^\star \right) (x_t, \pi_t^\star(x_t))dx_t}_{L}. \tag{170}$$

We analyze each term individually.

Bound on the L term. Whenever $x_t \in \mathcal{L}_t$, Corollary B.3 bounds the misspecification error so that:

$$L = \int_{\mathcal{L}_t} w_t(x_t) \left(H - t + 1 - Q_t^\star(x_t, \pi_t^\star(x_t)) \right) dx_t \tag{171}$$

$$\geqslant \int_{\mathcal{L}_t} w_t(x_t)[-(H - t + 1)\epsilon]dx_t \geqslant -4H\epsilon \int_{\mathcal{L}_t} w_t(x_t)dx_t \tag{172}$$

Bound on the S term. In states where the Q function is linear, the decomposition from Lemma C.1 gives us:

$$S = \int_{\mathcal{S}_t} w_t(x_t) \left\{ \mathbb{E}_{x'|x_t, \pi_t^\star(x_t)}[(\widetilde{V}_{t+1,k} - V_{t+1}^\star)(x')] + \phi_t(x_t, \pi_t^\star(x_t))^\top \left(\widetilde{\eta}_{tk} + \widetilde{\xi}_{tk} + \widetilde{\lambda}_{tk}^\star \right) + \widetilde{m}_{tk}^\star(x_t, \pi_t^\star(x_t)) \right\} dx_t. \tag{173}$$

Since we condition on the good event, applying Definitions D.2 and D.4 we have

$$S \geqslant \int_{\mathcal{S}_t} w_t(x_t) \left\{ \mathbb{E}_{x'|x_t,\pi_t^\star(x_t)}[(\widetilde{V}_{t+1,k} - V_{t+1}^\star)(x')] + \phi_t(x_t, \pi_t^\star(x_t))^\top \widetilde{\xi}_{tk} - \sqrt{\nu_k(\delta)} \|\phi_t(x_t, \pi_t^\star(x_t))\|_{\Sigma_{tk}^{-1}} - 4\epsilon H) \right\} dx_t. \tag{174}$$

Bound on the M term. This term interpolates between the values we would get out of the linearity of the representation and the default values. Define q^1 and q^2 to be the coefficient of the linear interpolation (see (109)), then:

$$M = \int_{\mathcal{M}_t} w_t(x_t) \left(q^1 \widetilde{Q}_t(x_t, \pi_t^\star(x_t)) + q^2(H - t + 1) - Q^\star(x_t, \pi_t^\star(x_t)) \right) dx_t \tag{175}$$

$$= \int_{\mathcal{M}_t} w_t(x_t) \left(q^1 \underbrace{(\widetilde{Q}_t - Q^\star)(x_t, \pi_t^\star(x_t))}_{\text{as in } S} + q^2 \underbrace{((H - t + 1) - Q^\star(x_t, \pi_t^\star(x_t)))}_{\text{as in } L} + \underbrace{(q^1 + q^2 - 1)}_{=0} Q^\star(x_t, \pi_t^\star(x_t)) \right) dx_t \tag{176}$$

$$\geqslant \int_{\mathcal{M}_t} w_t(x_t) \frac{\|\phi_t(x_t, \pi^\star(x_t))\|_{\Sigma_{tk}^{-1}} - \alpha_L}{\alpha_U - \alpha_L} \left\{ \mathbb{E}_{x'|x_t,\pi_t^\star(x_t)}[(\widetilde{V}_{t+1,k} - V_{t+1}^\star)(x')] + \phi_t(x_t, \pi_t^\star(x_t))^\top \widetilde{\xi}_{tk} \right. \tag{177}$$

$$\left. - \sqrt{\nu_k(\delta)} \|\phi_t(x_t, \pi_t^\star(x_t))\|_{\Sigma_{tk}^{-1}} - 4\epsilon H) \right\} dx_t - 4H\epsilon \int_{\mathcal{M}_t} w_t(x_t) \frac{\alpha_U - \|\phi_t(x_t, \pi^\star(x_t))\|_{\Sigma_{tk}^{-1}}}{\alpha_U - \alpha_L} dx_t \tag{178}$$

$$= \int_{\mathcal{M}_t} w_t(x_t) \frac{\|\phi_t(x_t, \pi^\star(x_t))\|_{\Sigma_{tk}^{-1}} - \alpha_L}{\alpha_U - \alpha_L} \left\{ \mathbb{E}_{x'|x_t,\pi_t^\star(x_t)}[(\widetilde{V}_{t+1,k} - V_{t+1}^\star)(x')] + \phi_t(x_t, \pi_t^\star(x_t))^\top \widetilde{\xi}_{tk} \right. \tag{179}$$

$$\left. - \sqrt{\nu_k(\delta)} \|\phi_t(x_t, \pi_t^\star(x_t))\|_{\Sigma_{tk}^{-1}} \right) \right\} dx_t - 4H\epsilon \int_{\mathcal{M}_t} w_t(x_t) dx_t \tag{180}$$

$$\tag{181}$$

Conclusion. Together, the bounds on S, M, L we have obtained can be combined (also with the definition of $\mathring{w}$) to obtain:

$$S + M + L \geqslant \int_{\mathcal{S}} \mathring{w}_t(x_t) \left\{ \mathbb{E}_{x'|x_t,\pi_t^\star(x_t)}[(\widetilde{V}_{t+1,k} - V_{t+1}^\star)(x')] + \phi_t(x_t, \pi_t^\star(x_t))^\top \widetilde{\xi}_{tk} - \sqrt{\nu_k(\delta)} \|\phi_t(x_t, \pi_t^\star(x_t))\|_{\Sigma_{tk}^{-1}} \right\} dx_t \tag{182}$$

$$- 4H\epsilon \int_{\mathcal{S}} w_t(x_t) dx_t \tag{183}$$

Using the statement made in (160) we must have:

$$-4H\epsilon \int_{\mathcal{S}} w_t(x_t) dx_t \geqslant -4H\epsilon. \tag{184}$$

In the end, using the definition in (155), we have obtained:

$$\int_{\mathcal{S}} w_t(x_t) \left(\widetilde{V}_t - V_t^\star \right)(x_t) dx_t \geqslant S + M + L \tag{185}$$

$$\geqslant \int_{\mathcal{S}} w_{t+1}(x_{t+1}) \left(\widetilde{V}_{t+1,k} - V_{t+1}^\star \right)(x_{t+1}) dx_t \tag{186}$$

$$+ \int_{\mathcal{S}} w_t(x_t) \left(\phi_t(x_t, \pi_t^\star(x_t))^\top \widetilde{\xi}_{tk} - \sqrt{\nu_k(\delta)} \|\phi_t(x_t, \pi_t^\star(x_t))\|_{\Sigma_{tk}^{-1}} \right) dx_t \tag{187}$$

$$- 4H\epsilon \tag{188}$$

Induction and plugging in the definition of $\phi_t^\star$ concludes the proof. $\qquad\square$

Lemma F.2 (Optimism). *For any episode k if $\Phi(-1)/2\delta > 1$ and $\epsilon < \frac{1}{10H}$:*

$$\mathbf{P}\left(\widetilde{V}_1(s_{1k}) - V_1^\star(s_{1k}) + 4H^2\epsilon \geq 0 \mid s_{1k}, \mathcal{H}_k\right) \geq \Phi(-1)/2 \tag{189}$$

Proof. All events in this lemma are conditioned on $s_{1k}, \mathcal{H}_k$ so that the only random variables are $\widetilde{\xi}_{tk}$ for $t \in [H]$. Consider the probability of being optimistic at the beginning of episode k, and call this event $\widetilde{\mathcal{O}}_k$:

$$\widetilde{\mathcal{O}}_k = \left\{\left(\widetilde{V}_{1k} - V_1^\star\right)(s_{1k}) \geq -4H^2\epsilon\right\}. \tag{190}$$

For $\Phi(-1)/2\delta > 1$ and with $\epsilon < \frac{1}{10H}$, by elementary probability and using Lemma E.6 to bound the probability of the good event:

$$\mathbf{P}(\widetilde{\mathcal{O}}_k) = 1 - \mathbf{P}(\widetilde{\mathcal{O}}_k^c) = 1 - \mathbf{P}(\widetilde{\mathcal{O}}_k^c \cap \widetilde{\mathcal{G}}_k) - \mathbf{P}(\widetilde{\mathcal{O}}_k^c \cap \widetilde{\mathcal{G}}_k^c) \geq 1 - \mathbf{P}(\widetilde{\mathcal{O}}_k^c \cap \widetilde{\mathcal{G}}_k) - \mathbf{P}(\widetilde{\mathcal{G}}_k^c) \tag{191}$$

$$\geq 1 - \mathbf{P}(\widetilde{\mathcal{O}}_k^c \cap \widetilde{\mathcal{G}}_k) - \delta. \tag{192}$$

Notice that under $\mathcal{G}_k$, Lemma F.1 allows us to deduce that

$$\left(\widetilde{V}_1 - V_1^\star\right)(s_{1k}) \geq \sum_{t=1}^{H}[(\phi_t^\star)^\top \widetilde{\xi}_{tk} - \sqrt{\nu_k(\delta)}\|\phi_t^\star\|_{\Sigma_{tk}^{-1}}] - 4H^2\epsilon. \tag{193}$$

So, defining

$$\mathcal{W}_k \stackrel{def}{=} \left\{\left(\widetilde{V}_1 - V_1^\star\right)(s_{1k}) \geq \sum_{t=1}^{H}[(\phi_t^\star)^\top \widetilde{\xi}_{tk} - \sqrt{\nu_k(\delta)}\|\phi_t^\star\|_{\Sigma_{tk}^{-1}}] - 4H^2\epsilon\right\}, \tag{194}$$

we have that

$$\mathbf{P}\left(\widetilde{\mathcal{O}}_k^c \cap \widetilde{\mathcal{G}}_k\right) \leq \mathbf{P}\left(\widetilde{\mathcal{O}}_k^c \cap \mathcal{W}_k\right). \tag{195}$$

Along with equation (191) we get:

$$\mathbf{P}(\widetilde{\mathcal{O}}_k) \geq 1 - \mathbf{P}\left(\widetilde{\mathcal{O}}_k^c \cap \mathcal{W}_k\right) - \delta. \tag{196}$$

Now, define the event that the random walk is positive in episode k:

$$\mathcal{P}_k = \left\{\sum_{t=1}^{H}[(\phi_t^\star)^\top \widetilde{\xi}_{tk} - \sqrt{\nu_k(\delta)}\|\phi_t^\star\|_{\Sigma_{tk}^{-1}}] \geq 0\right\} \tag{197}$$

Now note that chaining the inequalities from the definitions of $\overline{\mathcal{O}}_k$ and $\mathcal{W}_k$ we can see that

$$\widetilde{\mathcal{O}}_k \cap \mathcal{W}_k \subseteq \mathcal{P}_k^c. \tag{198}$$

Thus we have

$$\mathbf{P}(\widetilde{\mathcal{O}}_k) \geq 1 - \mathbf{P}\left(\mathcal{P}_k^c\right) - \delta \geq \mathbf{P}\left(\mathcal{P}_k\right) - \delta. \tag{199}$$

Recall that by the definition in the algorithm:

$$\widetilde{\xi}_{tk} \sim \mathcal{N}(0, H\nu_k(\delta)\Sigma_{tk}^{-1}). \tag{200}$$

Now, since we have conditioned on $\mathcal{H}_k$ and s_{1k}, by Lemma F.1 we have that $\phi_t^\star$ is non-random and thus by properties of the normal distribution:

$$(\phi_t^\star)^\top \widetilde{\xi}_{tk} \sim \mathcal{N}\left(0, H\nu_k(\delta)\|\phi_t^\star\|_{\Sigma_{tk}^{-1}}^2\right) \tag{201}$$

and

$$\sum_{t=1}^{H} (\phi_t^{\star})^{\top} \widetilde{\xi}_{tk} \sim \mathcal{N}\left(0, H\nu_k(\delta) \sum_{t=1}^{H} \|\phi_t^{\star}\|_{\Sigma_{tk}^{-1}}^2 \right). \tag{202}$$

Applying Cauchy-Schwarz we get that

$$\sum_{t=1}^{H} \sqrt{\nu_k(\delta)} \|\phi_t^{\star}\|_{\Sigma_{tk}^{-1}} \leqslant \sqrt{H\nu_k(\delta)} \left(\sum_{t=1}^{H} \|\phi_t^{\star}\|_{\Sigma_{tk}^{-1}}^2 \right)^{1/2}, \tag{203}$$

which is the standard deviation of the above random variable. Thus, we can conclude that

$$\mathbf{P}(\mathcal{P}_k) \geqslant \mathbf{P}\left(\sum_{t=1}^{H} (\phi_t^{\star})^{\top} \widetilde{\xi}_{tk} \geqslant \sqrt{H\nu_k(\delta)} \left(\sum_{t=1}^{H} \|\phi_t^{\star}\|_{\Sigma_{tk}^{-1}}^2 \right)^{1/2}\right) \geqslant \Phi(-1). \tag{204}$$

Plugging this in to (199) and ensuring that we choose $\delta < \Phi(-1)/2$ we get the result. $\qquad\square$

G Regret Bound

In this section we prove the main regret bound. This is split into two parts: one for the estimation error of each $\overline{V}_{tk}$ compared to $V_t^{\pi_k}$ and one for the pessimism of $\overline{V}_{tk}$ compared to $V_t^\star$.

G.1 Main theorem statement

Theorem G.1 (Main Result: High Probability Regret Bound for RLSVI with Approximately Linear Rewards and Low-Rank Transitions). *Under Assumption B.1 with $\Phi(-1)/2 > \delta > 0$ and $\lambda = 1$ and choosing $\alpha_L, \alpha_U, \sigma^2 = H\nu_k(\delta)$ as defined in Section D and letting $T = HK$, with probability at least $1 - \delta$ for OPT-RLSVI jointly for all episodes K:*

$$\text{REGRET}(K) \stackrel{def}{=} \sum_{k=1}^{K} \left(V_1^\star - V_1^{\pi_k} \right)(s_{1k}) = \tilde{O}\left(\sqrt{\gamma_K(\delta)}\sqrt{dHT} + \frac{H^2 d}{\alpha_L^2} + \epsilon HT \right). \tag{205}$$

Proof. We have the following decomposition:

$$\text{REGRET}(K) \stackrel{def}{=} \sum_{k=1}^{K} \left(V_1^\star - V_1^{\pi_k} \right)(s_{1k}) = \sum_{k=1}^{K} \left(V_1^\star - \overline{V}_{1k} \right)(s_{1k}) + \sum_{k=1}^{K} \left(\overline{V}_{1k} - V_1^{\pi_k} \right)(s_{1k}). \tag{206}$$

Taking a union bound over the results of Lemma G.3 and Lemma G.4 yields the result. $\qquad\square$

Corollary G.2 (High Probability Regret Bound for RLSVI with Approximately Linear Rewards and Low-Rank Transitions). *Under Assumption B.1 and if additionally $L_\phi = \tilde{O}(1)$, and $L_\psi, L_r = \tilde{O}(d)$, then with probability at least $1 - \delta$ for OPT-RLSVI it holds that:*

$$\text{REGRET}(K) \stackrel{def}{=} \sum_{k=1}^{K} \left(V_1^\star - V_1^{\pi_k} \right)(s_{1k}) = \tilde{O}\left(H^2 d^2 \sqrt{T} + H^5 d^4 + \epsilon dH(1 + \epsilon dH^2)T \right). \tag{207}$$

Proof. Recall from Definition D.2 we have that

$$\sqrt{\gamma_K(\delta)} = \tilde{O}((Hd)^{3/2} + \sqrt{Hd\lambda}L_\phi(3HL_\psi + L_r) + \epsilon\sqrt{dHT}) = \tilde{O}((Hd)^{3/2} + \epsilon\sqrt{dHT}) \tag{208}$$

And combining with Definition D.3 we have

$$\frac{1}{\alpha_L^2} = \tilde{O}((Hd)^3 + \epsilon^2 dHT) \tag{209}$$

Plugging these values into Theorem G.1 we get with probability at least $1 - \delta$ that

$$\text{REGRET}(K) = \tilde{O}\left(((Hd)^{3/2} + \epsilon\sqrt{dHT})\sqrt{dHT} + (H^2 d)((Hd)^3 + \epsilon^2 dHT) + \epsilon HT \right) \tag{210}$$

$$= \tilde{O}\left(H^2 d^2 \sqrt{T} + \epsilon dHT + H^5 d^4 + \epsilon^2 d^2 H^3 T + \epsilon HT \right) \tag{211}$$

$$= \tilde{O}\left(H^2 d^2 \sqrt{T} + H^5 d^4 + \epsilon dH(1 + \epsilon dH^2)T \right). \tag{212}$$

$\qquad\square$

G.2 Bounding the estimation error

Lemma G.3 (Bound on Estimation). *For any $\delta > 0$, it holds with probability at least $1 - \delta/2$ that:*

$$\sum_{k=1}^{K} \left(\overline{V}_{1k} - V_1^{\pi_k} \right)(s_{1k}) = \tilde{O}\left((\sqrt{\nu_K(\delta)} + \sqrt{\gamma_K(\delta)})H\sqrt{d}\sqrt{K} + H^2 K\epsilon + \frac{H^2 d}{\alpha_L^2} \right) \tag{213}$$

Proof. The proof proceeds by induction over $t \in [H]$ followed by some algebra to get the bound. Denote by G_k the event that $\overline{\mathcal{G}}_\ell$ (see Def. D.4) holds for all $\ell \leq k$, so that G_k is measurable with respect to $\overline{\mathcal{H}}_k$.

Consider a generic timestep t: we split into two cases. Either 1) we have $\|\phi_{tk}\|_{\Sigma_{tk}^{-1}} \leq \alpha_L$ which we will call $\mathcal{S}_{tk}$ or 2) we have $\|\phi_{tk}\|_{\Sigma_{tk}^{-1}} > \alpha_L$ which we will call $\mathcal{S}_{tk}^c$. Under $\mathcal{S}_{tk}$ the Q function is linear (see Eq. 109 or Def. 1), and under $\mathcal{S}_{tk}^c$ we can upper bound the value function difference by H in the worst case under G_k. Thus we have

$$\left(\overline{V}_{tk} - V_t^{\pi_k}\right)(s_{tk}) \mathbb{1}\{G_k\} = \mathbb{1}\{G_k\}\left(\left(\overline{V}_{tk} - V_t^{\pi_k}\right)(s_{tk})\mathbb{1}\{\mathcal{S}_{tk}\} + \left(\overline{V}_{tk} - V_t^{\pi_k}\right)(s_{tk})\mathbb{1}\{\mathcal{S}_{tk}^c\}\right) \quad (214)$$

$$= \mathbb{1}\{G_k\}\left(\left(\overline{V}_{tk}(s_{tk}) - Q_t^{\pi_k}(s_{tk}, a_{tk})\right)\mathbb{1}\{\mathcal{S}_{tk}\} + \left(\overline{V}_{tk} - V_t^{\pi_k}\right)(s_{tk})\mathbb{1}\{\mathcal{S}_{tk}^c\}\right) \quad (215)$$

$$\leq \mathbb{1}\{G_k\}\left(\underbrace{\left(\phi_{tk}^\top \overline{\theta}_{tk} - Q_t^{\pi_k}(s_{tk}, a_{tk})\right)\mathbb{1}\{\mathcal{S}_{tk}\}}_{S} + \underbrace{H\mathbb{1}\{\mathcal{S}_{tk}^c\}}_{S^c}\right). \quad (216)$$

We focus on the first term, term S. Applying Lemma C.1 we have

$$\phi_{tk}^\top \overline{\theta}_{tk} - Q_t^{\pi_k}(s_{tk}, a_{tk}) = \mathbb{E}_{s'|s,a}\left[\left(\overline{V}_{t+1,k} - V_{t+1}^{\pi_k}\right)(s')\right] + \phi_{tk}^\top(\overline{\eta}_{tk} + \overline{\xi}_{tk} + \overline{\lambda}_{tk}^{\pi_k}) + \overline{m}_{tk}^{\pi_k}(s, a). \quad (217)$$

And under $\overline{\mathcal{G}}_k$ we can bound this by

$$\phi_{tk}^\top \overline{\theta}_{tk} - Q_t^{\pi_k}(s_{tk}, a_{tk}) \leq \mathbb{E}_{s'|s_{tk}, a_{tk}}\left[\left(\overline{V}_{t+1,k} - V_{t+1}^{\pi_k}\right)(s')\right] + \left(\sqrt{\nu_k(\delta)} + \sqrt{\gamma_k(\delta)}\right)\|\phi_{tk}\|_{\Sigma_{tk}^{-1}} + 4H\epsilon. \quad (218)$$

Then we can define

$$\dot{\zeta}_{tk} \stackrel{def}{=} \mathbb{1}\{G_k\}\mathbb{1}\{\mathcal{S}_{tk}\}\left(\mathbb{E}_{s'|s_{tk}, a_{tk}}\left[\left(\overline{V}_{t+1,k} - V_{t+1}^{\pi_k}\right)(s')\right] - \left(\overline{V}_{t+1,k} - V_{t+1}^{\pi_k}\right)(s_{t+1,k})\right). \quad (219)$$

Note that due to the indicator of G_k we have that each $|\dot{\zeta}_{tk}| \leq 2H$ a.s. and $\mathbb{E}[\dot{\zeta}_{tk}|\overline{\mathcal{H}}_k \cup \mathcal{H}_{tk}] = 0$. Then $(\dot{\zeta}_{tk}, \overline{\mathcal{H}}_k \cup \mathcal{H}_{tk})_{t,k}$ is an MDS. So, applying Azuma-Hoeffding we have with probability at least $1 - \delta/4$ that $\sum_{k=1}^K \sum_{t=1}^H \dot{\zeta}_{tk} = \widetilde{O}(H\sqrt{T})$.

With this definition,

$$\mathbb{1}\{G_k\}S \leq \mathbb{1}\{G_k\}\mathbb{1}\{\mathcal{S}_{tk}\}\left(\left(\overline{V}_{t+1,k} - V_{t+1}^{\pi_k}\right)(s_{t+1,k}) + \left(\sqrt{\nu_k(\delta)} + \sqrt{\gamma_k(\delta)}\right)\|\phi_{tk}\|_{\Sigma_{tk}^{-1}} + 4H\epsilon\right) + \dot{\zeta}_{tk}. \quad (220)$$

Combining it all we have

$$\mathbb{1}\{G_k\}\left(\overline{V}_{tk} - V_t^{\pi_k}\right)(s_{tk}) \leq \mathbb{1}\{G_k\}\Bigg[\left(\overline{V}_{t+1,k} - V_{t+1}^{\pi_k}\right)(s_{t+1,k})\mathbb{1}\{\mathcal{S}_{tk}\} \quad (221)$$

$$+ \left(\left(\sqrt{\nu_k(\delta)} + \sqrt{\gamma_k(\delta)}\right)\|\phi_{tk}\|_{\Sigma_{tk}^{-1}} + 4H\epsilon\right)\mathbb{1}\{\mathcal{S}_{tk}\} + H\mathbb{1}\{\mathcal{S}_{tk}^c\}\Bigg] + \dot{\zeta}_{tk}. \quad (222)$$

And induction gives us

$$\mathbb{1}\{G_k\}\left(\overline{V}_{1k} - V_1^{\pi_k}\right)(s_{tk}) \leq \mathbb{1}\{G_k\}\sum_{t=1}^H\Bigg[\left(\left(\sqrt{\nu_k(\delta)} + \sqrt{\gamma_k(\delta)}\right)\|\phi_{tk}\|_{\Sigma_{tk}^{-1}} + 4H\epsilon\right)\left(\Pi_{\tau=1}^t\mathbb{1}\{\mathcal{S}_{\tau k}\}\right) \quad (223)$$

$$+ H\left(\Pi_{\tau=1}^{t-1}\mathbb{1}\{\mathcal{S}_{\tau k}\}\right)\mathbb{1}\{\mathcal{S}_{tk}^c\}\Bigg] + \sum_{t=1}^H \dot{\zeta}_{tk} \quad (224)$$

Now we can sum over k to attain a bound on the estimation error term of the regret. We will split this in three terms: when all $\mathcal{S}_{tk}$ occur, when some $\mathcal{S}_{tk}^c$ occurs, and the martingale difference terms. We can bound the dominant term by exchanging order of summation, pulling out constants, applying Cauchy-Schwarz, and finally

applying Lemma I.2 to get

$$\sum_{k=1}^{K} \mathbb{1}\{G_k\} \sum_{t=1}^{H} \left((\sqrt{\nu_k(\delta)} + \sqrt{\gamma_k(\delta)}) \|\phi_{tk}\|_{\Sigma_{tk}^{-1}} + 4H\epsilon \right) \left(\Pi_{\tau=1}^{t} \mathbb{1}\{\mathcal{S}_{\tau k}\} \right) \tag{225}$$

$$\leqslant (\sqrt{\nu_K(\delta)} + \sqrt{\gamma_K(\delta)}) \sum_{t=1}^{H} \sum_{k=1}^{K} \mathbb{1}\{G_k\} \|\phi_{tk}\|_{\Sigma_{tk}^{-1}} \left(\Pi_{\tau=1}^{t} \mathbb{1}\{\mathcal{S}_{\tau k}\} \right) + 4H^2 K\epsilon \tag{226}$$

$$\leqslant (\sqrt{\nu_K(\delta)} + \sqrt{\gamma_K(\delta)}) \sum_{t=1}^{H} \sqrt{K} \left(\sum_{k=1}^{K} \|\phi_{tk}\|_{\Sigma_{tk}^{-1}}^2 \left(\Pi_{\tau=1}^{t} \mathbb{1}\{\mathcal{S}_{\tau k}\} \right) \right)^{1/2} + 4H^2 K\epsilon \tag{227}$$

$$\leqslant (\sqrt{\nu_K(\delta)} + \sqrt{\gamma_K(\delta)}) \sum_{t=1}^{H} \sqrt{K} \left(\sum_{k=1}^{K} \min\{1, \|\phi_{tk}\|_{\Sigma_{tk}^{-1}}^2\} \right)^{1/2} + 4H^2 K\epsilon \tag{228}$$

$$\leqslant (\sqrt{\nu_K(\delta)} + \sqrt{\gamma_K(\delta)}) H\sqrt{K} \tilde{O}(\sqrt{d}) + 4H^2 K\epsilon \tag{229}$$

To get the conclusion we need to show that the desired bound holds with high probability. Note that if $\epsilon > \frac{1}{10H}$ the bound we are trying to prove is trivially true since it is larger than T. So, assuming $\epsilon < \frac{1}{10H}$ and applying Lemma E.6 we get that $\bigcap_{k \in [K]} G_k = \bigcap_{k \in [K]} \overline{\mathcal{G}}_k$ occurs with probability at least $1 - \delta/4$. Taking a union bound we see that with probability at least $1 - \delta/2$ both the G_k and the bound on the sum of the $\dot{\zeta}_{tk}$ hold. Adding the lower order term bound from Lemma G.5 gives the desired result. $\qquad\square$

G.3 Bounding the pessimism

Lemma G.4 (Bound on Pessimism). *For any* $\Phi(-1)/2 > \delta > 0$ *it holds with probability at least* $1 - \delta/2$ *that:*

$$\sum_{k=1}^{K} \left(V_{1k}^{\star} - \overline{V}_{1k} \right)(s_{1k}) = \tilde{O}\left((\sqrt{\nu_K(\delta)} + \sqrt{\gamma_K(\delta)}) H\sqrt{d}\sqrt{K} + H^2 K\epsilon + \frac{H^2 d}{\alpha_L^2} \right). \tag{230}$$

Proof. In this section we bound the pessimism term by connecting it to the probability of the algorithm being optimistic and the concentration terms. Essentially, we construct an upper bound on $V^{\star}$ and a lower bound on $\overline{V}_{1k}$ and show that they cannot be too different from each other.

As in the previous proof, we will use indicator functions of a good event. But, in this proof we will not just have the $\overline{\xi}$ pseudonoise variables but also $\widetilde{\xi}$ and $\underline{\xi}$ (defined later in the proof). These variables have good events $\widetilde{\mathcal{G}}_k, \underline{\mathcal{G}}_k$ defined per episode analogous to $\overline{\mathcal{G}}_k$ (see Def. D.4). Accordingly we will now denote by G_k the event that $\overline{\mathcal{G}}_\ell \cap \widetilde{\mathcal{G}}_\ell \cap \underline{\mathcal{G}}_\ell$ holds for all $\ell \leqslant k$, so that G_k is measurable with respect to $\overline{\mathcal{H}}_k$. Note that by Lemma E.6 and a union bound over the three pseudonoises we have that $\bigcap_{k \in [K]} G_k = \bigcap_{k \in [K]} \overline{\mathcal{G}}_k$ occurs with probability at least $1 - \delta'$ for any $\delta' > 0$.

First we construct the lower bound. Let the ξ_{tk}'s be vectors in $\mathbb{R}^d$ for $t = 1, \ldots, H$, and let V_{tk}^{ξ} be the value function obtained by running the Least Square Value Iteration procedure in Algorithm 1 backward with the non-random ξ_{tk} (see definition below) in place of $\overline{\xi}_{tk}$. Consider the following minimization program:

$$\min_{\{\xi_{tk}\}_{t=1,\ldots,H}} V_{1k}^{\xi}(s_{1k})$$
$$\|\xi_{tk}\|_{\Sigma_{tk}} \leqslant \sqrt{\gamma_k(\delta)}, \quad \forall t \in [H] \tag{231}$$

Notice that the constraint condition on the ξ variables is equivalent to the one on the $\overline{\xi}$ in the definition of $\mathcal{G}_{tk}^{\overline{\xi}}$ in Definition D.4, but with ξ_{tk} replacing the $\overline{\xi}_{tk}$. We denote with $\{\underline{\xi}_{tk}\}_{t=1,\ldots,H}$ a minimizer of the above expression and with $\underline{V}_{1k}(s_{1k})$ the minimum of the optimization program (the minimum exists because $V_{1k}^{\xi}(s_{1k})$ is a continuous function of the ξ which are defined on a compact set). Importantly, under $\overline{\mathcal{G}}_k$ we get that

$$\underline{V}_{1k}(s_{1k}) \leqslant \overline{V}_{1k}(s_{1k}) \tag{232}$$

because $\{\overline{\xi}_{tk}\}_{t=1,\ldots,H}$ is a feasible solution of the optimization and $V_{tk}^{\overline{\xi}}(s_{1k}) = \overline{V}_{tk}(s_{1k})$.

Next, we want to get an upper bound. Consider drawing an independent and identically distributed copy $\widetilde{\xi}_{tk}$ of the $\overline{\xi}_{tk}$'s and run the least square procedure backward to get a new value function $\widetilde{V}_{tk}$ (for $t \in [H]$) and action-value function $\widetilde{Q}_{tk}$. Define as $\widetilde{\mathcal{O}}_k$ the event that $\widetilde{V}_{1k}(s_{1k})$ is optimistic in the k-th episode. Applying Lemma F.2 with $\Phi(-1)/2 > \delta > 0$ and $\epsilon \leqslant \frac{1}{10H}$,

$$\mathbf{P}\left(\widetilde{\mathcal{O}}_k\right) = \mathbf{P}\left(\{\widetilde{V}_{1k}(s_{1k}) \geqslant V_{1k}^{\star}(s_{1k}) - 4H^2\epsilon\}\right) \geqslant \Phi(-1)/2. \tag{233}$$

Next using this definition of optimism we can write:

$$\left(V_{1k}^{\star} - \overline{V}_{1k}\right)(s_{1k})\mathbb{1}\left\{\overline{\mathcal{G}}_k\right\} \leqslant \mathbb{E}_{\widetilde{\xi}|\widetilde{\mathcal{O}}_k}\left[\left(\widetilde{V}_{1k} - \overline{V}_{1k}\right)(s_{1k})\right]\mathbb{1}\left\{\overline{\mathcal{G}}_k\right\} + 4H^2\epsilon \tag{234}$$

$$\leqslant \mathbb{E}_{\widetilde{\xi}|\widetilde{\mathcal{O}}_k}\left[\left(\widetilde{V}_{1k} - \underline{V}_{1k}\right)(s_{1k})\right]\mathbb{1}\left\{\overline{\mathcal{G}}_k\right\} + 4H^2\epsilon. \tag{235}$$

where the expectations are over the $\widetilde{\xi}$'s, conditioned on the event $\widetilde{\mathcal{O}}_k$. The second bound follows from Equation (232).

At this point we can use the law of total expectation under $\widetilde{\mathcal{G}}_k$:

$$\mathbb{E}_{\widetilde{\xi}}\left[\left(\widetilde{V}_{1k} - \underline{V}_{1k}\right)(s_{1k})\right] = \mathbb{E}_{\widetilde{\xi}|\widetilde{\mathcal{O}}_k}\left[\left(\widetilde{V}_{1k} - \underline{V}_{1k}\right)(s_{1k})\right]\mathbf{P}(\widetilde{\mathcal{O}}_k) + \underbrace{\mathbb{E}_{\widetilde{\xi}|\widetilde{\mathcal{O}}_k^c}\left[\left(\widetilde{V}_{1k} - \underline{V}_{1k}\right)(s_{1k})\right]\mathbf{P}(\widetilde{\mathcal{O}}_k^c)}_{\geqslant 0} \tag{236}$$

$$\geqslant \mathbb{E}_{\widetilde{\xi}|\widetilde{\mathcal{O}}_k}\left[\left(\widetilde{V}_{1k} - \underline{V}_{1k}\right)(s_{1k})\right]\mathbf{P}(\widetilde{\mathcal{O}}_k). \tag{237}$$

The lower bound again follows because $\{\widetilde{\xi}_{tk}\}_{t=1,\dots,H}$ is a feasible solution of (231), so the neglected term is positive. Chaining the above with (233) and (234) and using the definition of G_k (i.e., $G_k \implies \overline{\mathcal{G}}_k$):

$$\mathbb{1}\left\{G_k\right\}\left(V_{1k}^{\star} - \overline{V}_{1k}\right)(s_{1k}) \leqslant \mathbb{1}\left\{G_k\right\}\frac{2}{\Phi(-1)}\mathbb{E}_{\widetilde{\xi}}\left[\left(\widetilde{V}_{1k} - \underline{V}_{1k}\right)(s_{1k})\right] + 4H^2\epsilon \tag{238}$$

$$= \mathbb{1}\left\{G_k\right\}\frac{2}{\Phi(-1)}\left(\overline{V}_{1k} - \underline{V}_{1k}\right)(s_{1k}) + \ddot{\zeta}_k + 4H^2\epsilon \tag{239}$$

$$= \mathbb{1}\left\{G_k\right\}\frac{2}{\Phi(-1)}\left(\overline{V}_{1k} - V_1^{\pi_k} + V_1^{\pi_k} - \underline{V}_{1k}\right)(s_{1k}) + \ddot{\zeta}_k + 4H^2\epsilon \tag{240}$$

where we define

$$\ddot{\zeta}_k \stackrel{def}{=} \mathbb{1}\left\{G_k\right\}\frac{2}{\Phi(-1)}\left(\mathbb{E}_{\widetilde{\xi}}\left[\widetilde{V}_{1k}(s_{1k})\right] - \overline{V}_{1k}(s_{1k})\right) \tag{241}$$

and note that since the $\overline{\xi}_{tk}$ and $\widetilde{\xi}_{tk}$ are iid, so are $\widetilde{V}_{1k}$ and $\overline{V}_{1k}$. Then $(\ddot{\zeta}_k, \mathcal{H}_{k-1})_k$ is an MDS and due to the indicator function each term is bounded in absolute value by 2H. So, applying Azuma-Hoeffding we have with probability at least $1 - \delta'$ that $\sum_{k=1}^K \ddot{\zeta}_{tk} = \widetilde{O}(H\sqrt{K})$.

Now we decompose

$$\mathbb{1}\left\{G_k\right\}\left(\overline{V}_{1k} - V_1^{\pi_k} + V_1^{\pi_k} - \underline{V}_{1k}\right)(s_{1k}) = \mathbb{1}\left\{G_k\right\}\left(\overline{V}_{1k} - V_1^{\pi_k}\right)(s_{1k}) + \mathbb{1}\left\{G_k\right\}\left(V_1^{\pi_k} - \underline{V}_{1k}\right)(s_{1k}) \tag{242}$$

The first term is the estimation error that we bounded in Lemma G.3.

For the second term, we can derive the same bound, but require a slightly modified proof. As before, we set up the recursion by considering a generic timestep t and splitting into cases, bounding the difference by H on $\mathcal{S}_{tk}^c$ (see definition in Lem. G.3):

$$\mathbb{1}\left\{G_k\right\}\left(V_t^{\pi_k} - \underline{V}_{tk}\right)(s_{tk}) \leqslant \mathbb{1}\left\{G_k\right\}\left(\left(V_t^{\pi_k} - \underline{V}_{tk}\right)(s_{tk})\mathbb{1}\left\{\mathcal{S}_{tk}\right\} + H\mathbb{1}\left\{\mathcal{S}_{tk}^c\right\}\right) \tag{243}$$

Now consider the term where $\|\phi_{tk}\|_{\Sigma_{tk}^{-1}} \leqslant \alpha_L$ holds. First note that since a_{tk} is the action that maximizes $\overline{Q}_{tk}$,

$$\left(V_t^{\pi_k} - \underline{V}_{tk}\right)(s_{tk}) = Q_t^{\pi_k}(s_{tk}, a_{tk}) - \underline{V}_{tk}(s_{tk}) \leqslant \left(Q_t^{\pi_k} - \underline{Q}_{tk}\right)(s_{tk}, a_{tk}) = Q_t^{\pi_k}(s_{tk}, a_{tk}) - \phi_{tk}^{\top}\underline{\theta}_{tk}. \tag{244}$$

Applying Lemma C.1 we see that this is

$$Q_t^{\pi_k}(s_{tk}, a_{tk}) - \phi_{tk}^\top \underline{\theta}_{tk} = -\mathbb{E}_{s'|s_{tk},a_{tk}}[(\underline{V}_{t+1,k} - V_{t+1}^{\pi_k})(s')] - \phi_{tk}^\top(\underline{\eta}_{tk} + \underline{\xi}_{tk} + \underline{\lambda}_{tk}^{\pi_k}) - \underline{m}_{tk}^{\pi_k}(s_{tk}, a_{tk}). \tag{245}$$

And we can define

$$\dddot{\zeta}_{tk} \stackrel{def}{=} \mathbb{1}\{G_k\}\left(-\mathbb{E}_{s'|s_{tk},a_{tk}}[(\underline{V}_{t+1,k} - V_{t+1}^{\pi_k})(s')] + (\underline{V}_{t+1,k} - V_{t+1}^{\pi_k})(s_{t+1,k})\right) \tag{246}$$

Then $(\dddot{\zeta}_{tk}, \overline{\mathcal{H}}_k \cup \mathcal{H}_{tk})_{t,k}$ is an MDS and due to the indicator function each term is bounded in absolute value by 2H. So, applying Azuma-Hoeffding we have with probability at least $1 - \delta'$ that $\sum_{k=1}^K \sum_{t=1}^H \dddot{\zeta}_{tk} = \tilde{O}(H\sqrt{T})$

So that, as in Lemma G.3, induction gives us

$$\mathbb{1}\{G_k\}(V_1^{\pi_k} - \underline{V}_{1k})(s_{1k}) \leqslant \mathbb{1}\{G_k\}\sum_{t=1}^H\left[\left((\sqrt{\nu_k(\delta)} + \sqrt{\gamma_k(\delta)})\|\phi_{tk}\|_{\Sigma_{tk}^{-1}} + 4H\epsilon\right)(\Pi_{\tau=1}^t \mathbb{1}\{\mathcal{S}_{\tau k}\})\right. \tag{247}$$

$$\left. + H\left(\Pi_{\tau=1}^{t-1}\mathbb{1}\{\mathcal{S}_{\tau k}\}\right)\mathbb{1}\{\mathcal{S}_{tk}^c\}\right] + \sum_{t=1}^H \dddot{\zeta}_{tk} \tag{248}$$

Summing over k, this can be bounded as in Lemma G.3. To conclude, summing the bound from (240) over k and applying the same arguments as Lemma G.3 to both value function differences gives us that

$$\sum_{k=1}^K \mathbb{1}\{G_k\}(V_{1k}^\star - \overline{V}_{1k})(s_{1k}) \leqslant \frac{4}{\Phi(-1)}\tilde{O}\left((\sqrt{\nu_K(\delta)} + \sqrt{\gamma_K(\delta)})H\sqrt{d}\sqrt{K} + H^2 K\epsilon + \frac{H^2 d}{\alpha_L^2}\right) + \tilde{O}(H\sqrt{K}) + 4HT\epsilon \tag{249}$$

so that consolidating terms gives us the desired bound. Notice that if $\epsilon \geqslant \frac{1}{10H}$ the result trivially holds.

To conclude we just need to set $\delta' = \delta/6$ and take a union bound over the two applications of Azuma-Hoeffding and the intersection of the G_k we get the result with probability $1 - \delta/2$ as desired. □

G.4 Bounding the warmup

Lemma G.5 (Warmup Bound).

$$\sum_{k=1}^K \sum_{t=1}^H H\mathbb{1}\{\mathcal{S}_{tk}^c\} \stackrel{def}{=} \sum_{k=1}^K \sum_{t=1}^H H\mathbb{1}\left\{\|\phi_{tk}\|_{\Sigma_{tk}^{-1}} > \alpha_L\right\} = \tilde{O}\left(\frac{H^2 d}{\alpha_L^2}\right). \tag{250}$$

Proof.

$$\sum_{k=1}^K \sum_{t=1}^H H\mathbb{1}\left\{\|\phi_{tk}\|_{\Sigma_{tk}^{-1}} > \alpha_L\right\} = H\sum_{k=1}^K \sum_{t=1}^H \mathbb{1}\left\{\frac{\|\phi_{tk}\|_{\Sigma_{tk}^{-1}}}{\alpha_L} > 1\right\} \tag{251}$$

$$= H\sum_{k=1}^K \sum_{t=1}^H \mathbb{1}\left\{\frac{\|\phi_{tk}\|_{\Sigma_{tk}^{-1}}^2}{\alpha_L^2} > 1\right\} \tag{252}$$

$$\leqslant H\sum_{k=1}^K \sum_{t=1}^H \min\left\{1, \frac{\|\phi_{tk}\|_{\Sigma_{tk}^{-1}}^2}{\alpha_L^2}\right\} \tag{253}$$

$$\stackrel{(a)}{\leqslant} \frac{H}{\alpha_L^2}\sum_{t=1}^H \sum_{k=1}^K \min\{1, \|\phi_{tk}\|_{\Sigma_{tk}^{-1}}^2\} \tag{254}$$

$$\stackrel{(b)}{\leqslant} \frac{H^2}{\alpha_L^2}\tilde{O}(d) = \tilde{O}\left(\frac{H^2 d}{\alpha_L^2}\right) \tag{255}$$

Where (a) holds since $1/\alpha^2 > 1$ by the following reasoning. Let $x > 1$ and consider two cases: if $y < 1/x$ then $\min\{1, xy\} = xy = x\min\{1, y\}$ and if $y \geqslant 1/x$ then $\min\{1, xy\} = 1 \leqslant x \leqslant x\min\{1, y\}$. Finally, (b) is due to Lemma I.2. □

Note that Lemma C.1 is derived for $\overline{\theta}_{tk}$, but we can derive an equivalent expression for $\underline{\theta}_{tk}$

H   Computational Complexity

Now we take a look at the computational complexity of the algorithm.

Proposition H.1 (Computational Complexity of OPT-RLSVI in finite action spaces)**.** *Let A be the number of actions available at every timestep. Then* OPT-RLSVI *can be implemented in space $O(d^2H + dAHK)$ and time $O(d^2AHK^2)$.*

Proof. In terms of computational complexity, a naive implementation of OPT-RLSVI requires $O(d^2)$ elementary operations to compute $\|\phi_{t+1,i}\|_{\Sigma_{t+1,k}^{-1}}$ to assess which decision rule to use in definition 1. This must be done for all next-state action-value functions at the experienced successors states. If the action space is finite with cardinality A then the maximization over action to compute the value function $\overline{V}_{t+1,k}(s_{t+1,i})$ at the next timestep for the k experienced successor states $s_{t+1,1}, \ldots, s_{t+1,k}$ would take $O(d^2AK)$ total work per timestep. A further $O(d^3)$ is needed to compute the inverse of Σ_{tk} to solve the least square system of equation, but this can be brought down to $O(d^2)$ using the usual Sherman-Morrison rank one update formula. All this must be done at every timestep of the least-square value iteration procedure, which must run every episode, giving a final runtime $O(d^2AHK^2)$.

As for the memory, one can store the K features $\phi_t(s_{tk}, a)$ for all A actions, timestep H and episode K using $O(dAHK)$ memory, in addition to the inverse of the Σ_{tk} matrices ($O(d^2H)$ space) and the scalar rewards ($O(KH)$ space). □

I Technical Lemmas

Lemma I.1 (Self-normalized process). *(Abbasi-Yadkori et al., 2011) Let $\{x_i\}_{i=1}^{\infty}$ be a real valued stochastic process sequence over the filtration $\{\mathcal{F}_i\}_{i=1}^{\infty}$. Let x_i be conditionally B-subgaussian given $\mathcal{F}_{i-1}$. Let $\{\phi_i\}_{i=1}^{\infty}$ with $\phi_i \in \mathcal{F}_{i-1}$ be a stochastic process in $\mathbb{R}^d$ with each $\|\phi_i\| \leqslant L_\phi$. Define $\Sigma_i = \lambda I + \sum_{j=1}^{i-1} \phi_i \phi_i^\top$. Then for any $\delta > 0$ and all $i \geqslant 0$, with probability at least $1 - \delta$*

$$\left\| \sum_{i=1}^{k-1} \phi_i x_i \right\|_{\Sigma_k^{-1}}^2 \leqslant 2B^2 \log\left(\frac{\det(\Sigma_i)^{1/2} \det(\lambda I)^{-1/2}}{\delta} \right) \leqslant 2B^2 \left(d \log\left(\frac{\lambda + kL_\phi^2}{\lambda} \right) + \log(1/\delta) \right) \tag{256}$$

Lemma I.2 (Sum of features). *(Abbasi-Yadkori et al., 2011, Lemma 11) Using the notation defined above,*

$$\sum_{i=1}^{k} \min\{1, \|\phi_i\|_{\Sigma_i^{-1}}^2\} \leqslant 2d \log\left(\frac{\lambda + kL_\phi^2}{\lambda} \right) \tag{257}$$

Lemma I.3 (Sum of features in final norm). *(Jin et al., 2019, Lemma D.1)*

$$\sum_{i=1}^{k-1} \|\phi_i\|_{\Sigma_k^{-1}}^2 \leqslant d \tag{258}$$

Lemma I.4 (Gaussian concentration). *(Abeille et al., 2017, Appendix A) Let $\overline{\xi}_{tk} \sim \mathcal{N}(0, H\nu_k(\delta)\Sigma_{tk}^{-1})$. For any $\delta > 0$, with probability $1 - \delta$*

$$\|\overline{\xi}_{tk}\|_{\Sigma_{tk}} \leqslant c\sqrt{Hd\nu_k(\delta)\log(d/\delta)} \stackrel{def}{=} \sqrt{\gamma_k(\delta)} \tag{259}$$

for some absolute constant c

Lemma I.5 (Covering numbers). *(Pollard, 1990, Section 4) A euclidean ball of radius B in $\mathbb{R}^d$ has ε-covering number at most $(3B/\varepsilon)^d$.*

Lemma I.6 (Simplifying the log term). *With $\lambda \geqslant 1$, we can choose c_1 so that*

$$\sqrt{\beta_k(\delta)} \geqslant 8Hd\left(\log\left(\frac{kL_\phi^2 + \lambda}{\lambda} \right) + \log\left(3(2H\sqrt{kd/\lambda} + \sqrt{\gamma_k(\delta)/\lambda} + 1/\lambda)/\left(\frac{\alpha_U - \alpha_L}{8kL_\phi^2 H^2} \right)^2 \right) + \log(1/\delta) \right)^{1/2} \tag{260}$$

Proof. Recall that

$$\sqrt{\beta_k(\delta)} \stackrel{def}{=} c_1 Hd \sqrt{\log\left(\frac{Hdk \max(1, L_\phi) \max(1, L_\psi) \max(1, L_r)\lambda}{\delta} \right)} \tag{261}$$

Using $\lambda \geqslant$ and expanding the definitions of terms on the RHS of the statement we can bound it by

$$\leqslant 8Hd\left(\log\left(\frac{(kL_\phi^2 + \lambda)3(2H\sqrt{kd} + \sqrt{\gamma_k(\delta)} + 1)64k^2L_\phi^4 H^4}{\delta(\alpha_U - \alpha_L)^2} \right) \right)^{1/2} \tag{262}$$

$$\leqslant 8Hd\left(\log\left(\frac{(kL_\phi^2 + \lambda)(H\sqrt{kd} + \sqrt{\gamma_k(\delta)})64k^2L_\phi^4 H^2(\gamma_k(\delta))}{\delta\lambda} \right) \right)^{1/2} \tag{263}$$

$$\leqslant 8Hd\left(\log\left(\frac{(kL_\phi^2 + \lambda)(H\sqrt{kd})k^2L_\phi^4 H^2(c_2^2 dH(\sqrt{\beta_k(\delta)} + \sqrt{\lambda}L_\phi(3HL_\psi + L_r) + 4\epsilon H\sqrt{dk})^2 \log(d/\delta))^{3/2}}{\delta} \right) \right)^{1/2} \tag{264}$$

Bounding the $\sqrt{\beta_k(\delta)}$ by $c_1 Hd(Hdk \max(1, L_\phi) \max(1, L_\psi) \max(1, L_r)\lambda)/\delta$ this gives us a large polynomial in $k, H, d, \lambda, \max(1, L_\phi), \max(1, L_\psi), \max(1, L_r), 1/\delta$. We bound this by $c(kHd\lambda \max(1, L_\phi) \max(1, L_\psi) \max(1, L_r)/\delta)^{c'}$ for some c, c', and taking the log to move the exponent into the constant gives the existence of some c_1 to define $\beta_k(\delta)$. $\qquad\square$