Bridging Exploration and General Function Approximation in Reinforcement Learning: Provably Efficient Kernel and Neural Value Iterations

Zhuoran Yang* Chi Jin* Zhaoran Wang† Mengdi Wang* Michael I. Jordan‡

November 10, 2020

Abstract

[1] Reinforcement learning (RL) algorithms combined with modern function approximators such as kernel functions and deep neural networks have achieved significant empirical successes in large-scale application problems with a massive number of states. From a theoretical perspective, however, RL with functional approximation poses a fundamental challenge to developing algorithms with provable computational and statistical efficiency, due to the need to take into consideration both the exploration-exploitation tradeoff that is inherent in RL and the bias-variance tradeoff that is innate in statistical estimation. To address such a challenge, focusing on the episodic setting where the action-value functions are represented by a kernel function or over-parametrized neural network, we propose the first provable RL algorithm with both polynomial runtime and sample complexity, without additional assumptions on the data-generating model. In particular, for both the kernel and neural settings, we prove that an optimistic modification of the least-squares value iteration algorithm incurs an $\widetilde{\mathcal{O}}(\delta_{\mathcal{F}} H^2 \sqrt{T})$ regret, where $\delta_{\mathcal{F}}$ characterizes the intrinsic complexity of the function class $\mathcal{F}$, H is the length of each episode, and T is the total number of episodes. Our regret bounds are independent of the number of states and therefore even allows it to diverge, which exhibits the benefit of function approximation.

1 Introduction

Reinforcement learning (RL) algorithms combined with modern function approximators such as kernel functions and deep neural networks have achieved tremendous empirical successes in a variety of application problems (e.g., Duan et al., 2016; Silver et al., 2016, 2017; Wang et al., 2018; Vinyals et al., 2019). However, when practically powerful function approximators are employed,

*Princeton University, {zy6,chij,mendiw}@princeton.edu

†Northwestern University, zhaoranwang@gmail.com

‡University of California, Berkeley, jordan@cs.berkeley.edu

[1]Part of this work was done while the Z. Yang, C. Jin, Z. Wang, and M. Wang are visiting the Simons Institute for the Theory of Computing.

designing RL algorithms with provable computational and statistical efficiency remains a challenging task. In particular, from a theoretical perspective, function approximation brings statistical estimation into the scope of RL, resulting in the need to balance the bias-variance tradeoff that is innate in statistical estimation and the exploration-exploitation tradeoff that is inherent in RL at the same time. As a result, most existing provably efficient RL algorithms are only applicable only to the tabular setting (see, e.g., Jaksch et al., 2010; Osband et al., 2014; Azar et al., 2017; Jin et al., 2018; Osband et al., 2018; Russo, 2019) where both the state and action spaces are discrete and the value function can be represented as a table.

Provably efficient exploration under the function approximation setting has only been studied recently with most of the existing work consider a (generalized) linear model (Yang and Wang, 2019b,a; Jin et al., 2019; Cai et al., 2019a; Zanette et al., 2020a; Wang et al., 2019). However, their algorithms and analyses stem from those of upper confidence bound (UCB) or Thompson sampling for linear contextual bandits (Bubeck and Cesa-Bianchi, 2012; Lattimore and Szepesvári, 2018) and thus are confined to the linear setting. Moreover, the linear assumption is rather rigid and rarely satisfied in practice. When such a model is misspecified, their methods would suffer from a linear regret. In addition, going beyond the linear model, various recent works have proposed provably sample efficient algorithms with general function approximation. However, these methods either are computationally intractable or (Krishnamurthy et al., 2016; Jiang et al., 2017; Dann et al., 2018; Dong et al., 2019) or hinge on rather strong assumptions on the transition model (Wen and Van Roy, 2017; Du et al., 2019b). Thus, the following question is left open:

> Can we design an RL algorithm with powerful function approximators such as neural networks or kernel functions that provably achieves both computational and statistical efficiency?

In this work, we provide an affirmative answer to such a question. Specifically, we focus on the setting of episodic Markov decision process (MDP) where the value function is represented by either a kernel function or an overparametrized neural network. For both cases, we propose the first provable efficient RL algorithm with both polynomial runtime and sample complexity, without any additional assumptions on the data-generating model. Our algorithm is an optimistic modification of the least-squares value iteration algorithm (LSVI) (Bradtke and Barto, 1996) — a classical batch RL algorithm — where we add an additional UCB bonus term to each iterate of the LSVI updates to promote efficient exploration. Specifically, when using a kernel function, each LSVI update becomes a kernel ridge regression, and the bonus term is motivated from that proposed for kernelized contextual bandits (Srinivas et al., 2009; Valko et al., 2013; Chowdhury and Gopalan, 2017). Moreover, for the neural network setting, motivated by the NeuralUCB algorithm for contextual bandits (Zhou et al., 2019), we construct a UCB bonus from the tangent features of the neural network and each LSVI updated is solved via projected gradient descent with a finite number of iterations. For both these settings, by adding the UCB bonus, we ensure value functions constructed by the algorithm are always optimistic in the sense that they all serve as upper bounds of the optimal value function. Furthermore, for both the kernel and neural settings, we prove that the proposed algorithm incurs an $\widetilde{\mathcal{O}}(\delta_{\mathcal{F}} H^2 \sqrt{T})$ regret, H is the length of each episode, T is the total number of episodes, and $\delta_{\mathcal{F}}$ quantifies the intrinsic complexity of the function class $\mathcal{F}$. Specifically, as we will show in §4, $\delta_{\mathcal{F}}$ is determined by the interplay between the ℓ_∞-covering number of the value function class adopted by the algorithm and the effective dimension of function class $\mathcal{F}$. (See Table 1 for a summary.) A key feature of our regret bounds is that they depends on the complexity of the state space only through $\delta_{\mathcal{F}}$ and thus allow the number of states to be very large or even divergent, which exhibits the benefit of function approximation in terms of the sample efficiency. To the best of our knowledge, we establish the provably efficient reinforcement learning algorithm with kernel and neural network function approximations for the first time.

function class $\mathcal{F}$	regret bound
general RKHS $\mathcal{H}$	$H^2 \cdot \sqrt{d_{\text{eff}} \cdot [d_{\text{eff}} + \log N_\infty(\epsilon^*)]} \cdot \sqrt{T}$
γ-finite spectrum	$H^2 \cdot \sqrt{\gamma^3 T} \cdot \log(\gamma T H)$
γ-exponential decay	$H^2 \cdot \sqrt{(\log T)^{3/\gamma} \cdot T} \cdot \log(TH)$
γ-polynomial decay	$H^2 \cdot T^{\kappa^* + \xi^* + 1/2} \cdot [\log(TH)]^{3/2}$
overparameterized neural network	$H^2 \cdot \sqrt{d_{\text{eff}} \cdot [d_{\text{eff}} + \log N_\infty(\epsilon^*)]} \cdot \sqrt{T} + \mathrm{poly}(T, H) \cdot m^{-1/12}$

Table 1: Summary of the main results. Here H is the length of each episode, T is the number of episodes in total, and $2m$ is the number of neurons of the overparameterized networks in the neural setting. For an RKHS $\mathcal{H}$ in general, d_{eff} denotes the effective dimension of $\mathcal{H}$ and $N_\infty(\epsilon^*)$ is the ℓ_∞-covering number of the value function class, where $\epsilon^* = H/T$. Moreover, to obtain concrete bounds, we also apply the general result to RKHS with various eigenvalue decay conditions. Here γ is a positive integer in the case of γ-finite spectrum and is a positive integer in the subsequent two cases. In addition, κ^* and ξ^* are defined in (4.9), which are constants that depend on d, the input dimension, and the parameter γ. Finally, in the last case we present the regret bound for the neural setting in general, where d_{eff} is the effective dimension of the neural tangent kernel (NTK) induced by the overparameterized neural network with $2m$ neurons and $\mathrm{poly}(T, H)$ is a polynomial of T and H. Such a general regret can be made in concrete form depending on the spectral property of the NTK.

Related Work. Our work belongs to the vast literature on establishing provably efficient RL methods without having access to a generative model or a explorative behavioral policy. The tabular setting is well studied the existing works. See, e.g., Jaksch et al. (2010); Osband et al. (2014); Azar et al. (2017); Dann et al. (2017); Strehl et al. (2006); Jin et al. (2018); Russo (2019) and the references therein. It is shown in Azar et al. (2017); Jin et al. (2018) that any RL algorithm necessarily incurs a $\Omega(\sqrt{SAT})$ regret under the tabular setting, where S and A are the cardinalities of the state and action spaces, respectively. Thus, the algorithms designed for the tabular setting cannot be directly applied to the function approximation setting where the number of states is gigantic. When function approximation is employed, Yang and Wang (2019a,b); Jin et al. (2019); Cai et al. (2019a); Zanette et al. (2020a); Wang et al. (2019); Ayoub et al. (2020); Zhou et al. (2020); Kakade et al. (2020) focus on the (generalized) linear setting where the value function (or the transition model) can be represented using a linear transform of a known feature mapping. Among these works, our work is most related to Jin et al. (2019). In particular, in our kernel setting, when kernel function has a finite rank, both our LSVI algorithm and the corresponding regret bound are reduced to the those established in Jin et al. (2019). However, their sample complexity or regret bounds all diverge when the dimension of the feature mapping goes to infinity and thus cannot be directly extended to the kernel setting. Another closely related work is Wang et al. (2020), which studies a similar optimistic LSVI algorithm for general function approximation. Their work focuses on value function classes with bounded eluder dimensions (Russo and Van Roy, 2013; Osband and Van Roy, 2014) and it is unclear whether their construction of the bonus function can be extended to the kernel or neural settings. Besides, Yang and Wang (2019b) also study a kernelized MDP model where the transition model can be directly estimated. Under a slightly more general model, Ayoub et al. (2020) recently propose an optimistic model-based algorithm via value-targeted regression, where the model class is allowed to be general functions with bounded eluder dimension. In another recent work, Kakade et al. (2020) study a nonlinear control problem where the system dynamics belongs to a known RKHS and can be directly estimated from the data. As opposed to these works, we do not pose an explicit assumption on the transition model and our proposed algorithm is model-free. Furthermore, regret or sample complexity results have also been studied beyond linear function approximation. However, these algorithms are either computational challenging (Krishnamurthy et al., 2016; Jiang et al., 2017; Dann et al., 2018; Dong et al., 2019) or require additional assumptions on the transition model that might be restrictive (Wen and Van Roy, 2013, 2017; Du et al., 2019b)

In addition, our work is also related to the literature on contextual bandits with kernel or (Srinivas et al., 2009; Krause and Ong, 2011; Srinivas et al., 2012; Valko et al., 2013; Chowdhury and Gopalan, 2017; Durand et al., 2018) neural network functions (Zhou et al., 2019), which are special cases of our episodic MDP with the episode length equal to one. The construction of our bonus function are adopted from these works. However, our reinforcement learning problem has temporal dependence caused by state transitions according to the Markov transition kernel, which is absent in bandit models. Specifically, the covering number $N_\infty(\epsilon^*)$ in Table 1 arises due to such an additional structure captures the fundamental challenge of temporally extended exploration in RL. When applying our algorithm to kernel contextual bandits, the regret bound reduces to $d_{\text{eff}} \cdot \sqrt{T}$ where d_{eff} is the effective dimension of the RKHS. Such a regret bound matches those in Srinivas et al. (2009); Chowdhury and Gopalan (2017).

Furthermore, our analysis of the optimistic LSVI algorithm is akin to the recent study of the optimization and generalization of over-parameterized neural networks via the framework of the neural tangent kernel (Jacot et al., 2018). Most of these works focus on the supervised learning (Daniely, 2017; Jacot et al., 2018; Wu et al., 2018; Du et al., 2018a,b; Allen-Zhu et al., 2018b,a; Zou et al., 2018; Chizat and Bach, 2018; Li and Liang, 2018; Arora et al., 2019; Cao and Gu, 2019a,b; Lee et al., 2019). In contrast, our algorithm incorporates an additional bonus term in the least-squares problem and thus requires novel analysis.

2 Background

In this section, we introduce the background of reinforcement learning, reproducing kernel Hilbert space (RKHS), and overparameterized neural networks.

2.1 Episodic Markov Decision Process

Throughout this work, we focus on an episodic MDP, which is denoted by $\mathrm{MDP}(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$. Here $\mathcal{S}$ and $\mathcal{A}$ are state and action spaces, respectively, integer $H > 0$ is the length of each episode, and $\mathbb{P} = \{\mathbb{P}_h\}_{h\in[H]}$ and $r = \{r_h\}_{h\in[H]}$ are the Markov transition kernel and the reward functions, respectively. Here we let $[n]$ denote the set $\{1, \ldots, n\}$ for all integer $n \geq 1$. Besides, we assume that $\mathcal{S}$ is a measurable space with possibly infinite number of elements and $\mathcal{A}$ is a finite set. Moreover, for each $h \in [H]$, $\mathbb{P}_h(\cdot \,|\, x, a)$ denotes the probability distribution of the next states if action a is taken at state $x \in \mathcal{S}$ in timestep $h \in [H]$, and $r_h \colon \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function at step h which is assumed to be deterministic for simplicity.

A policy π of an agent is a set of H functions $\pi = \{\pi_h\}_{h\in[H]}$ such that each $\pi_h(\cdot \,|\, x)$ is a probability distribution over $\mathcal{A}$. Here $\pi_h(a \,|\, x)$ is probability of the agent taking action a at state x at the h-th step in the episode. Furthermore, the agent interacts with the environment as follows. For any $t \geq 1$, at the beginning of the t-th episode, the agent determines a policy $\pi^t = \{\pi_h^t\}_{h\in[H]}$ while an initial state x_1^t is picked arbitrarily by the environment. Then, at each step $h \in [H]$, the agent observes the state $x_h^t \in \mathcal{S}$, picks an action $a_h^t \sim \pi_h^t(\cdot \,|\, x_h^t)$, and receives a reward $r_h(x_h^t, a_h^t)$. Moreover, the environment evolves into a new state x_{h+1}^t that is drawn from the probability measure $\mathbb{P}_h(\cdot \,|\, x_h^t, a_h^t)$. The episode terminates when the H-th step is reached and $r_H(x_H^t, a_H^t)$ is the last reward the agent receives.

The performance of the agent is captured by the value function. Specifically, for any policy π, and $h \in [H]$, we define the value function $V_h^\pi \colon \mathcal{S} \to \mathbb{R}$ as

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

Here we let $\mathbb{E}_\pi[\cdot]$ denote the expectation with respect to the randomness of the trajectory $\{(x_h, a_h)\}_{h=1}^H$ obtained by following policy π. Accordingly, we also define the action-value function $Q_h^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ as

$$Q_h^\pi(x, a) = \mathbb{E}_\pi\left[\sum_{h'=h}^{H} r_{h'}(x_{h'}, a_{h'}) \,\middle|\, x_h = x,\, a_h = a\right].$$

Moreover, let $\pi^\star$ denote the optimal policy which yields the optimal value function $V_h^\star(x) = \sup_\pi V_h^\pi(x)$ for all $x \in \mathcal{S}$ and $h \in [H]$. To simplify the notation, we denote

$$[\mathbb{P}_h V](x, a) := \mathbb{E}_{x' \sim \mathbb{P}_h(\cdot \,|\, x, a)}[V(x')]$$

for any measurable function $V \colon \mathcal{S} \to [0, H]$. Using this notation, the Bellman equation associated with a policy π becomes

$$Q_h^\pi(x, a) = (r_h + \mathbb{P}_h V_{h+1}^\pi)(x, a), \qquad V_h^\pi(x) = \langle Q_h^\pi(x, \cdot), \pi_h(\cdot \,|\, x)\rangle_{\mathcal{A}}, \qquad V_{H+1}^\pi(x) = 0. \tag{2.1}$$

Here we let $\langle \cdot, \cdot \rangle_{\mathcal{A}}$ denote the inner product over $\mathcal{A}$. Similarly, the Bellman optimality equation is given by

$$Q_h^\star(x,a) = (r_h + \mathbb{P}_h V_{h+1}^\star)(x,a), \qquad V_h^\star(x) = \max_{a \in \mathcal{A}} Q_h^\star(x,a), \qquad V_{H+1}^\star(x) = 0. \tag{2.2}$$

Thus, the optimal policy $\pi^\star$ is the greedy policy with respect to $\{Q_h^\star\}_{h \in [H]}$. Moreover, we define the Bellman optimality operator $\mathbb{T}_h^\star$ by letting

$$(\mathbb{T}_h^\star Q)(x,a) = r(x,a) + (\mathbb{P}_h V)(x,a) \qquad \text{for all } \ Q \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R},$$

where $V(x) = \max_{a \in \mathcal{A}} Q(x,a)$. By definition, the Bellman equation in (2.2) is equivalent to $Q_h^\star = \mathbb{T}_h^\star Q_{h+1}^\star$, $\forall h \in [H]$. The goal of the agent is to learn the optimal policy $\pi^\star$. For any policy π, the difference between V_1^π and $V_1^\star$ quantifies the sub-optimality of π. Thus, for a fixed integer $T > 0$, after playing for T episodes, the total (expected) regret (Bubeck and Cesa-Bianchi, 2012) of the agent is defined as

$$\text{Regret}(T) = \sum_{t=1}^{T} \big[V_1^\star(x_1^t) - V_1^{\pi^t}(x_1^t) \big],$$

where π^t is the policy executed in the t-th episode and x_1^t is the initial state.

2.2 Reproducing Kernel Hilbert Space

In the next section, we aim to estimate the optimal value function $Q_h^\star$ using functions in a reproducing kernel Hilbert space (RKHS) (Hofmann et al., 2008). To this end, hereafter, to simplify the notation, we let $z = (x,a)$ denote a state-action pair and denote $\mathcal{Z} = \mathcal{S} \times \mathcal{A}$. Without loss of generality, we regard $\mathcal{Z}$ as a compact subset of $\mathbb{R}^d$ where the dimension d is assumed fixed. This can be achieved if there exists a known embedding mapping $\psi_{\text{embed}} \colon \mathcal{Z} \to \mathbb{R}^d$ that pre-processes the input (x,a). Let $\mathcal{H}$ be an RKHS defined on $\mathcal{Z}$ with kernel function $K \colon \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$, which contains a family of functions defined on $\mathcal{Z}$. Let $\langle \cdot, \cdot \rangle_{\mathcal{H}} \colon \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ and $\| \cdot \|_{\mathcal{H}} \colon \mathcal{H} \to \mathbb{R}$ denote the inner product and RKHS norm on $\mathcal{H}$, respectively. Since $\mathcal{H}$ is an RKHS, there exists a feature mapping $\phi \colon \mathcal{Z} \to \mathcal{H}$ such that $f(z) = \langle f(\cdot), \phi(z) \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$ and all $z \in \mathcal{Z}$. Moreover, for any $x, y \in \mathcal{Z}$, we have $K(x,y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$. In this work, we assume that the kernel function K is uniformly bounded in the sense that $\sup_{z \in \mathcal{Z}} K(z,z) < \infty$. Without loss of generality, we assume that $\sup_{z \in \mathcal{Z}} K(z,z) \le 1$, which implies that $\|\phi(z)\|_{\mathcal{H}} \le 1$ for all $z \in \mathcal{Z}$.

Furthermore, let $\mathcal{L}^2(\mathcal{Z})$ be the space of square-integrable functions on $\mathcal{Z}$ with respect to the Lebesgue measure and let $\langle \cdot, \cdot \rangle_{\mathcal{L}^2}$ be the inner product on $\mathcal{L}^2(\mathcal{Z})$. The kernel function K induces a integral operator $T_K \colon \mathcal{L}^2(\mathcal{Z}) \to \mathcal{L}^2(\mathcal{Z})$ defined as

$$T_K f(z) = \int_{\mathcal{Z}} K(z,z') \cdot f(z') \, \mathrm{d}z', \qquad \forall f \in \mathcal{L}^2(\mathcal{Z}). \tag{2.3}$$

By Mercer's Theorem (Steinwart and Christmann, 2008), the integral operator T_K has countable and positive eigenvalues $\{\sigma_i\}_{i \ge 1}$ and the corresponding eigenfunctions $\{\psi_i\}_{i \ge 1}$ form an orthonormal basis of $\mathcal{L}^2(\mathcal{Z})$. Moreover, the kernel function admits a spectral expansion

$$K(z,z') = \sum_{i=1}^{\infty} \sigma_i \cdot \psi_i(z) \cdot \psi_j(z'). \tag{2.4}$$

Then, the RKHS $\mathcal{H}$ can be written as a subset of $\mathcal{L}^2(\mathcal{Z})$ as

$$\mathcal{H} = \bigg\{ f \in \mathcal{L}^2(\mathcal{Z}) \colon \sum_{i=1}^{\infty} \frac{\langle f, \psi_i \rangle_{\mathcal{L}^2}^2}{\sigma_i} < \infty \bigg\},$$

and the inner product of $\mathcal{H}$ can be written as

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{i=1}^{\infty} 1/\sigma_i \cdot \langle f, \psi_i \rangle_{\mathcal{L}^2} \cdot \langle g, \psi_i \rangle_{\mathcal{L}^2}, \qquad \text{for all} \quad f, g \in \mathcal{H}.$$

By such a construction, the scaled eigenfunctions $\{\sqrt{\sigma_i}\psi_i\}_{i \geq 1}$ form an orthogonal basis of RKHS $\mathcal{H}$ and the feature mapping $\phi(z) \in \mathcal{H}$ can be written as $\phi(z) = \sum_{i=1}^{\infty} \sigma_i \psi_i(z) \cdot \psi_i$ for any $z \in \mathcal{Z}$.

2.3 Overparameterized Neural Networks

In addition to RKHS, we also study the setting where the value functions are approximated by overparameterized neural networks. In the sequel, we define the class of neural networks that will be used in the algorithm.

Recall that we denote $\mathcal{Z} = \mathcal{S} \times \mathcal{A}$ and view it as a subset of $\mathbb{R}^d$. For neural networks, we further regard $\mathcal{Z}$ as a subset of the unit sphere in $\mathbb{R}^d$. That is, $\|z\|_2 = 1$ for all $z = (x, a) \in \mathcal{Z}$. A two-layer neural network $f(\cdot; b, W) \colon \mathcal{Z} \to \mathbb{R}$ with $2m$ neurons and weights (b, W) is defined as

$$f(z; b, W) = \frac{1}{\sqrt{2m}} \sum_{j=1}^{2m} b_j \cdot \mathrm{act}(W_j^\top z), \qquad \forall z \in \mathcal{Z}. \tag{2.5}$$

Here $\mathrm{act} \colon \mathbb{R} \to \mathbb{R}$ is the activation function, $b_j \in \mathbb{R}$ and $W_j \in \mathbb{R}^d$ for all $j \in [2m]$, and $b = (b_1, \ldots, b_{2m})^\top \in \mathbb{R}^{2m}$ and $W = (W_1, \ldots, W_{2m}) \in \mathbb{R}^{2dm}$. During training, we initialize (b, W) via the symmetric initialization scheme (Gao et al., 2019; Bai and Lee, 2019) as follows. For any $j \in [m]$, we set $b_j \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(\{-1, 1\})$ and $W_j \overset{\text{i.i.d.}}{\sim} N(0, I_d/d)$, where I_d is the identity matrix in $\mathbb{R}^d$. For any $j \in \{m+1, \ldots, 2m\}$, we set $b_j = -b_{j-m}$ and $W_j = W_{j-m}$. We remark that such an initialization implies that the initial neural network is a zero function, which is used only to simply the theoretical analysis. Besides, for ease of presentation, during training we fix b at its initial value and only optimize over W. Moreover, we denote $f(z; b, W)$ by $f(z; W)$ to simplify the notation.

Furthermore, we assume that the neural network in is overparameterized in the sense that the width $2m$ is much larger than the number of episodes T. Overparameterization is shown to be pivotal for neural training in both theory and practice (Neyshabur and Li, 2019; Allen-Zhu et al., 2018a; Arora et al., 2019). Under the such a regime, the dynamics of training neural networks are well captured by the framework of neural tangent kernel (NTK) (Jacot et al., 2018). Specifically, let $\varphi(\cdot; W) \colon \mathcal{Z} \to \mathbb{R}^{2md}$ be the gradient of $f(; W)$ with respect to W, which is given by

$$\varphi(z; W) = \nabla_W f(z; W) = \big(\nabla_{W_1} f(z; W), \ldots, \nabla_{W_{2m}} f(z; W)\big), \qquad \forall z \in \mathcal{Z}. \tag{2.6}$$

Let $W^{(0)}$ be the initial value of W. Condition on the realization of $W^{(0)}$, we define a kernel matrix $K_m \colon \mathcal{Z} \to \mathcal{Z}$ as

$$K_m(z, z') = \big\langle \varphi(z; W^{(0)}), \varphi(z'; W^{(0)}) \big\rangle, \qquad \forall (z, z') \in \mathcal{Z} \times \mathcal{Z}. \tag{2.7}$$

When m is sufficiently large, for all W that is in a neighborhood of $W^{(0)}$, it can be shown that $f(\cdot, W)$ is close to its linearization at $W^{(0)}$,

$$f(\cdot; W) \approx \widehat{f}(\cdot; W) = f(\cdot, W^{(0)}) + \big\langle \phi(\cdot; W^{(0)}), W - W^{(0)} \big\rangle = \big\langle \phi(\cdot; W^{(0)}), W - W^{(0)} \big\rangle. \tag{2.8}$$

The linearized function $\widehat{f}(\cdot; W)$ belongs to the RKHS with kernel K_m. Moreover, as m goes to infinity, due to random initialization, K_m converges to a kernel $K_{\text{ntk}} \colon \mathcal{Z} \times \mathcal{Z}$, dubbed as neural tangent kernel (NTK), which is given by

$$K_{\text{ntk}}(z, z') = \mathbb{E}\big[\text{act}'(w^\top z) \cdot \text{act}'(w^\top z') \cdot z^\top z'\big], \qquad (z, z') \in \mathcal{Z} \times \mathcal{Z}, \tag{2.9}$$

where act' is the derivative of the activation function, and the expectation in (2.9) is taken with respect to $w \sim N(0, I_d/d)$.

3 Optimistic Least-Squares Value Iteration Algorithms

In this section, we introduce the optimistic least-squares value iteration algorithm where the action-value functions are estimated using a class of functions defined on $\mathcal{Z} = \mathcal{S} \times \mathcal{A}$. The value iteration algorithm (Puterman, 2014; Sutton and Barto, 2018) is one of the most classical method in reinforcement learning, which finds $\{Q_h^\star\}_{h \in [H]}$ by applying the Bellman equation in (2.2) recursively. Specifically, value iteration constructs a sequence of action-value functions $\{Q_h\}_{h\in[H]}$ via

$$\begin{aligned} Q_h(x, a) &\leftarrow (\mathbb{T}_h^\star Q_{h+1}) = \big[r_h + \mathbb{P}_h V_{h+1}\big](x, a), \\ V_{h+1}(x) &\leftarrow \max_{a' \in \mathcal{A}} Q_{h+1}(x, a'), \qquad \forall (x, a) \in \mathcal{S} \times \mathcal{A}, \forall h \in [H], \end{aligned} \tag{3.1}$$

where Q_{H+1} is set to be the zero function. However, this algorithm is impractical to implement in real-world RL problems due to the following two reasons: (i) the transition kernel $\mathbb{P}_h$ is unknown and (ii) we can neither iterate over all state-action pairs nor store a table of size $|\mathcal{S} \times \mathcal{A}|$ when the number of states is large. To tackle these challenges, the least-squares value iteration (Bradtke and Barto, 1996; Osband et al., 2014) algorithm implements the update in (3.1) approximately by solving a least-squares regression problem based on historical data, which consists of the trajectories generated by the RL agent in previous episodes. Specifically, let $\mathcal{F}$ be a function class. Before the beginning of the t-th episode, we have observed $t-1$ transition tuples $\{(x_h^\tau, a_h^\tau, x_{h+1}^\tau)\}_{\tau \in [n]}$. Then, for estimating $Q_h^\star$, LSVI proposes to replace (3.1) with a least-squares regression problem

$$\widehat{Q}_h^t \leftarrow \operatorname*{minimize}_{f \in \mathcal{F}} \bigg\{ \sum_{\tau=1}^{t-1} \big[r_h(x_h^\tau, a_h^\tau) + V_{h+1}^t(x_{h+1}^\tau) - f(x_h^\tau, a_h^\tau)\big]^2 + \text{pen}(f) \bigg\}, \tag{3.2}$$

where $\text{pen}(f)$ is a regularization term. Moreover, to foster exploration, following the principle of optimism in the face of uncertainty (Sutton and Barto, 2018), we further incorporate a bonus function $b_h^t \colon \mathcal{Z} \to \mathbb{R}$ and define

$$Q_h^t(\cdot, \cdot) = \min\big\{\widehat{Q}_h^t(\cdot, \cdot) + \beta \cdot b_h^t(\cdot, \cdot), H - h + 1\big\}^+, \qquad V_h^t(\cdot) = \max_{a \in \mathcal{A}} Q_h^t(\cdot, a), \tag{3.3}$$

where $\beta > 0$ is a parameter and $\min\{\cdot, H-h+1\}^+$ denotes the truncation to the interval $[0, H-h-1]$. Here we truncate the value function to $[0, H-h+1]$ as each reward function is bounded in $[0, 1]$. Then, in the t-the episode, we let π^t be the greedy policy with respect to $\{Q_h^t\}_{h\in[H]}$ and execute π^t. Hence, combining (3.2) and (3.3) yields the optimistic least-squares value iteration algorithm, whose details are given in Algorithm 1.

Algorithm 1 Optimistic Least-Squares Value Iteration with Function Approximation

Input: Function class $\mathcal{F}$, penalty function $\mathrm{pen}(\cdot)$, and parameter β.
for episode $t = 1, \ldots, T$ **do**
  Receive the initial state x_1^t.
  Set V_{H+1}^t as the zero function.
  for step $h = H, \ldots, 1$ **do**
    Obtain Q_h^t and V_h^t according to (3.2) and (3.3).
  end for
  for step $h = 1, \ldots, H$ **do**
    Take action $a_h^t \leftarrow \mathrm{argmax}_{a\in\mathcal{A}} Q_h^t(x_h^t, a)$.
    Observe the reward $r_h(x_h^t, a_h^t)$ and the next state x_{h+1}^t.
  end for
end for

We note that the both the bonus function b_h^t in (3.3) and the penalty function in (3.2) relies on the choice of function class $\mathcal{F}$. The optimistic LSVI in Algorithm 1 is only implementable when $\mathcal{F}$ is specified. For instance, when $\mathcal{F}$ consists of functions of linear the form $\theta^\top\phi(z)$, where $\phi\colon \mathcal{Z} \to \mathbb{R}^d$ is a known feature mapping and $\theta \in \mathbb{R}^d$ is the parameter, we choose the ridge penalty $\|\theta\|_2^2$ in (3.2) and define $b_h^t(z)$ as $[\phi(z)^\top A_h^t \phi(z)]^{1/2}$ for some invertible matrix A_h^t. Then, Algorithm 1 recovers the LSVI-UCB algorithm studied in Jin et al. (2019), which further reduces to the tabular UCBVI algorithm (Azar et al., 2017) when ϕ is the canonical basis.

In the rest of this section, we instantiate the optimistic LSVI framework by setting $\mathcal{F}$ as an RKHS and the class of overparameterized neural networks.

3.1 The Kernel Setting

In the following, we consider the case where function class $\mathcal{F}$ is an RKHS $\mathcal{H}$ with kernel K. In this case, by setting $\mathrm{pen}(f)$ as the ridge penalty, (3.2) reduces to a kernel ridge regression problem. Besides, we define b_h^t in (3.3) as the UCB bonus function that also appears in kernelized contextual bandit (Srinivas et al., 2009; Valko et al., 2013; Chowdhury and Gopalan, 2017; Durand et al., 2018; Yang and Wang, 2019b; Sessa et al., 2019; Calandriello et al., 2019). With these two modifications, we obtain the <u>K</u>ernel <u>O</u>ptimistic Least-Squares <u>V</u>alue <u>I</u>teration (KOVI) algorithm, which is summarized in Algorithm 2.

Specifically, for each $t \in [T]$, before the beginning of the t-th episode, we first obtain value functions $\{Q_h^t\}_{h\in[H]}$ by solving a sequence of kernel ridge regressions with the data obtained from the previous $t-1$ episodes. In particular, we let Q_{H+1}^t be a zero function. For any $h \in [H]$, we replace (3.2) by a kernel ridge regression given by

$$\widehat{Q}_h^t \leftarrow \underset{f\in\mathcal{H}}{\text{minimize}} \sum_{\tau=1}^{t-1} \big[r_h(x_h^\tau, a_h^\tau) + V_{h+1}^t(x_{h+1}^\tau) - f(x_h^\tau, a_h^\tau)\big]^2 + \lambda \cdot \|f\|_{\mathcal{H}}^2, \tag{3.4}$$

where $\lambda > 0$ is the regularization parameter. Then, we obtain Q_h^t and V_h^t as in (3.3), where the bonus function b_h^t will be specified later. That is,

$$Q_h^t(s,a) = \min\big\{\widehat{Q}_h^t(s,a) + \beta \cdot b_h^t(s,a), H-h+1\big\}^+, \qquad V_h^t(s) = \max_a Q_h^t(s,a), \tag{3.5}$$

where $\beta > 0$ is a parameter.

The solution to (3.4) can be written in closed-form as follows. We define the response vector $y_h^t \in \mathbb{R}^{t-1}$ by letting its τ-th entry be

$$[y_h^t]_\tau = r_h(x_h^\tau, a_h^\tau) + V_{h+1}^t(x_{h+1}^\tau), \qquad \forall \tau \in [t-1]. \tag{3.6}$$

Recall that we denote $z = (x,a)$ and $\mathcal{Z} = \mathcal{S} \times \mathcal{A}$. Besides, based on the kernel function K of the RKHS, we define the Gram matrix $K_h^t \in \mathbb{R}^{(t-1)\times(t-1)}$ and function $k_h^t \colon \mathcal{Z} \to \mathbb{R}^{t-1}$ respectively as

$$K_h^t = [K(z_h^\tau, z_h^{\tau'})]_{\tau,\tau'\in[t-1]} \in \mathbb{R}^{(t-1)\times(t-1)}, \qquad k_h^t(z) = \big[K(z_h^1, z), \dots K(z_h^{t-1}, z)\big]^\top \in \mathbb{R}^{t-1}. \tag{3.7}$$

Then $\widehat{Q}_h^t$ in (3.4) can be written as $\widehat{Q}_h^t(z) = k_h^t(z)^\top \alpha_h^t$, where we define $\alpha_h^t = (K_h^t + \lambda \cdot I)^{-1} y_h^t$.

Using K_h^t and k_h^t defined in (3.7), the bonus function is defined as

$$b_h^t(x,a) = \lambda^{-1/2} \cdot \big[K(z,z) - k_h^t(z)^\top (K_h^t + \lambda I)^{-1} k_h^t(z)\big]^{1/2}, \tag{3.8}$$

which can be interpreted as the posterior variance of Gaussian process regression and characterizes the uncertainty of $\widehat{Q}_h^t$ (Rasmussen, 2003). Such a bonus term also appears in the literature on kernelized contextual bandits (Srinivas et al., 2009; Valko et al., 2013; Chowdhury and Gopalan, 2017; Durand et al., 2018; Yang and Wang, 2019b; Sessa et al., 2019; Calandriello et al., 2019) and is reduced to the UCB bonus proposed for linear bandits (Bubeck and Cesa-Bianchi, 2012; Lattimore and Szepesvári, 2018) when the feature mapping ϕ of the RKHS is finite-dimensional. In this case, KOVI reduces to the LSVI-UCB algorithm proposed in Jin et al. (2019) for linear value functions.

Furthermore, we remark that the bonus defined in (3.8) is called the UCB bonus because, when added by such a bonus function, Q_h^t defined in (3.5) serves as an upper bound of $Q_h^\star$ for all state-action pair. Intuitively, the target function of the kernel ridge regression in (3.4) is $\mathbb{T}_h^\star Q_{h+1}^t$. However, due to having limited data, the solution $\widehat{Q}_h^t$ has some estimation error, which is quantified b_h^t. Thus, when β is properly chosen, the bonus term triumphs the uncertainty of estimation, which yields that $Q_h^t \geq \mathbb{T}_h^\star Q_{h+1}^t$ elementwisely. Notice that $Q_{H+1}^t = Q_{H+1}^\star = 0$. The Bellman equation $Q_h^\star = \mathbb{T}_h^\star Q_{h+1}^\star$ directly implies that Q_h^t is an elementwise upper bound of $Q_h^\star$ for all $h \in [H]$. Our algorithm is called "optimistic value iteration" as the policy π^t is greedy with respect to $\{Q_h^t\}_{h\in[H]}$, which are upper bounds of the optimal value function. In other words, compared with the standard value iteration algorithm, we always over-estimate the value function. Such an optimistic approach is pivotal for the RL agent to perform efficient temporally extended exploration.

Algorithm 2 Kernelized Optimistic Least-Squares Value Iteration (KOVI)

Input: Parameters λ and β.
for episode $t = 1, \ldots, T$ **do**
    Receive the initial state x_1^t.
    Set V_{H+1}^t as the zero function.
    for step $h = H, \ldots, 1$ **do**
        Compute the response $y_h^t \in \mathbb{R}^{t-1}$, the Gram matrix $K_h^t \in \mathbb{R}^{(t-1)\times(t-1)}$, and function k_h^t as in (3.6) and (3.7), respectively.
        Compute
            $\alpha_h^t = (K_h^t + \lambda \cdot I)^{-1} y_h^t$,
            $b_h^t(\cdot,\cdot) = \lambda^{-1/2} \cdot \big[K(\cdot,\cdot;\cdot,\cdot) - k_h^t(\cdot,\cdot)^\top (K_h^t + \lambda I)^{-1} k_h^t(\cdot,\cdot)\big]^{1/2}$.
        Obtain value functions

$$Q_h^t(\cdot,\cdot) \leftarrow \min\{k_h^t(\cdot,\cdot)^\top \alpha_h^t + \beta \cdot b_h^t(\cdot,\cdot), H - h + 1\}^+, \qquad V_h^t(\cdot) = \max_a Q_h^t(\cdot, a).$$

    end for
    for step $h = 1, \ldots, H$ **do**
        Take action $a_h^t \leftarrow \operatorname{argmax}_{a\in\mathcal{A}} Q_h^t(x_h^t, a)$.
        Observe the reward $r_h(x_h^t, a_h^t)$ and the next state x_{h+1}^t.
    end for
end for

3.2 The Neural Setting

In this section, we estimate the value functions $\{Q_h^\star\}_{h\in[H]}$ using overparameterized neural networks. We aim to estimate each $Q_h^\star$ using a neural network given in (2.5), which is initialized via the symmetric initialization scheme (Gao et al., 2019; Bai and Lee, 2019) introduced in §2.3. Moreover, for simplicity, we assume that all the neural networks share the same initial weights, denoted by $(b^{(0)}, W^{(0)})$. Besides, we fix $b = b^{(0)}$ in (2.5) and only update the value of $W \in \mathbb{R}^{2md}$.

Under such a neural setting, we replace the least-squares regression in (3.2) by a nonlinear ridge regression. In particular, for any $(t,h) \in [T] \times [H]$, we define the loss function $L_h^t\colon \mathbb{R}^{2md} \to \mathbb{R}$ as

$$L_h^t(W) = \sum_{\tau=1}^{t-1} \big[r_h(x_h^\tau, a_h^\tau) + V_{h+1}^t(x_{h+1}^\tau) - f(x_h^\tau, a_h^\tau; W)\big]^2 + \lambda \cdot \big\|W - W^{(0)}\big\|_2^2, \tag{3.9}$$

where $\lambda > 0$ is the regularization parameter. Then we define $\widehat{Q}_h^t$ as

$$\widehat{Q}_h^t(\cdot,\cdot) = f\big(\cdot,\cdot;\widehat{W}_h^t\big), \qquad \text{where} \qquad \widehat{W}_h^t = \operatorname*{argmin}_{W\in\mathbb{R}^{2md}} L_h^t(W). \tag{3.10}$$

Here we assume that there is an optimization oracle that returns the global minimizer of the loss function L_h^t. It has been shown in a large body of literature that, when m is sufficiently large, with random initialization, simple optimization methods such as gradient descent provably find the global minimizer of the empirical loss function at a linear rate of convergence (Du et al., 2018b,a; Arora et al., 2019). Thus, such an optimization oracle can be realized by gradient descent with sufficiently large number of iterations and the computational cost of realizing such a oracle is polynomial in m, T, and H.

It remains to construct the bonus function b_h^t. Recall that we define $\varphi(\cdot; W) = \nabla_W f(\cdot; W)$ in (2.6). We define matrix $\Lambda_h^t \in \mathbb{R}^{2md \times 2md}$ as

$$\Lambda_h^t = \lambda \cdot I_{2md} + \sum_{\tau=1}^{t-1} \varphi\big(x_h^\tau, a_h^\tau; \widehat{W}_h^{\tau+1}\big)\varphi\big(x_h^\tau, a_h^\tau; \widehat{W}_h^{\tau+1}\big)^\top, \tag{3.11}$$

which can be recursively computed by letting

$$\Lambda_h^1 = \lambda \cdot I_{2md}, \qquad \Lambda_h^t = \Lambda_h^{t-1} + \varphi\big(x_h^{t-1}, a_h^{t-1}; \widehat{W}_h^t\big)\varphi\big(x_h^{t-1}, a_h^{t-1}; \widehat{W}_h^t\big)^\top, \qquad \forall t \geq 2.$$

Then the bonus function b_h^t is defined as

$$b_h^t(x, a) = \big[\varphi\big(x, a; \widehat{W}_h^t\big)^\top (\Lambda_h^t)^{-1} \varphi\big(x, a; \widehat{W}_h^t\big)\big]^{1/2}, \qquad \forall (x, a) \in \mathcal{S} \times \mathcal{A}. \tag{3.12}$$

Finally, we obtain the value functions Q_h^t and V_h^t via (3.5), with $\widehat{Q}_h^t$ and b_h^t defined in (3.10) and (3.12), respectively. By letting π^t be the greedy policy with respect to $\{Q_h^t\}_{h \in [H]}$, we obtain the <u>N</u>eural <u>O</u>ptimistic Least-Squares <u>V</u>alue <u>I</u>teration (NOVI) algorithm, whose details are stated in Algorithm 3 in §A.

The intuition of the bonus term in (3.12) can be understood via the connection between overparameterized neural networks and NTK. Specifically, when m is sufficiently large, it can be shown that each $\widehat{W}_h^t$ is not far from the initial value $W^{(0)}$. When this is the case, suppose we replace the neural tangent features $\{\varphi(\cdot; \widehat{W}_h^\tau)\}_{\tau \in [T]}$ in (3.11) and (3.12) by $\varphi(\cdot; W^{(0)})$, then b_h^t recovers the UCB bonus in linear contextual bandits and linear MDPs with feature mapping $\varphi(\cdot; W^{(0)})$ (Abbasi-Yadkori et al., 2011; Jin et al., 2019; Wang et al., 2019). Moreover, when m converges to infinity, it will become the UCB bonus defined in (3.8) for the RKHS setting with the kernel being K_{ntk}. Thus, when the neural networks are overparameterized, value functions $\{Q_h^t\}_{h \in [H]}$ are approximately elementwise upper bounds of the optimal value functions and thus we achieve optimism approximately.

4 Main Results

In this section, we prove that both KOVI and NOVI achieve $\mathcal{O}(\delta_{\mathcal{F}} H^2 \sqrt{T})$-regret bounds, where $\delta_{\mathcal{F}}$ characterizes the intrinsic complexity of the function class $\mathcal{F}$ used to approximate $\{Q_h^\star\}_{h \in [H]}$. We first consider the kernel setting as follows.

4.1 Regret of KOVI

Before presenting the theory, we first lay out a structural assumption for the kernel setting, which postulates that the Bellman operator maps any bounded value function to a bounded RKHS-norm ball.

Assumption 4.1. Let $R_Q > 0$ be a fixed constant. We define $\mathcal{Q}^\star = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \le R_Q H\}$. We assume that for any $h \in [H]$ and any $Q : \mathcal{S} \times \mathcal{A} \to [0, H]$, we have $\mathbb{T}_h^\star Q \in \mathcal{Q}^\star$.

Since $Q_h^\star$ is bounded by in $[0, H]$ for each all $h \in [H]$, Assumption 4.1 ensures the optimal value functions are contained in the RKHS-norm ball $\mathcal{Q}^\star$. Thus, there is no approximation bias when using functions in $\mathcal{H}$ to approximate $\{Q_h^\star\}_{h\in[H]}$. Moreover, it is shown in Du et al. (2019a) that only assuming $\{Q_h^\star\}_{h\in[H]} \subseteq \mathcal{Q}^\star$ is not sufficient for achieving a regret that is polynomial in H. Thus, we further assume that $\mathcal{Q}^\star$ contains the image of the Bellman operator. A sufficient condition for Assumption 4.1 to hold is that

$$r_h(\cdot,\cdot),\ \mathbb{P}_h(x' \,|\, \cdot,\cdot) \in \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \le 1\}, \qquad \forall h \in [H],\ \forall x' \in \mathcal{S}. \tag{4.1}$$

That is, both the reward function and the Markov transition kernel can be represented by functions in the unit ball of $\mathcal{H}$. When (4.1) holds, for any $V : \mathcal{S} \to [0, H]$, it holds that $r_h + \mathbb{P}_h V \in \mathcal{H}$ with its RKHS norm bounded by $H + 1$. Hence, Assumption 4.1 holds with $R_Q = 2$. Moreover, similar assumptions are also made in Yang and Wang (2019a,b); Jin et al. (2019); Zanette et al. (2020a,b); Wang et al. (2019) for (generalized) linear functions. Also see Du et al. (2019a); Van Roy and Dong (2019); Lattimore and Szepesvari (2019) for related discussions on the necessity of such an assumption.

Moreover, as $\mathcal{Q}^\star$ contains the image of the Bellman operator, the complexity of $\mathcal{H}$ plays an important role in the performance of KOVI. To characterize the intrinsic complexity of $\mathcal{F}$, we consider a notion of effective dimension named the maximal information gain (Srinivas et al., 2009), which is defined as

$$\Gamma_K(T, \lambda) = \sup_{\mathcal{D} \subseteq \mathcal{Z}} \big\{1/2 \cdot \operatorname{logdet}(I + K_{\mathcal{D}}/\lambda)\big\}, \tag{4.2}$$

where the supremum is taken over all $\mathcal{D} \subseteq \mathcal{Z}$ with $|\mathcal{D}| \le T$. Here in (4.2) $K_{\mathcal{D}}$ is the Gram matrix defined in the same way as in (3.7) based on $\mathcal{D}$, $\lambda > 0$ is a parameter, and the subscript K in Γ_K indicates the kernel K. The magnitude of $\Gamma_K(T, \lambda)$ relies on how fast the the eigenvalues $\mathcal{H}$ decay to zero and can be viewed as a proxy of the dimension of $\mathcal{H}$ when $\mathcal{H}$ is infinite-dimensional. In the special case where $\mathcal{H}$ is finite-rank, it holds that $\Gamma_K(T, \lambda) = \mathcal{O}(\gamma \cdot \log T)$ where γ is the rank of $\mathcal{H}$.

Furthermore, for any $h \in [H]$, note that each Q_h^t constructed by KOVI takes the form of

$$Q(z) = \min\Big\{Q_0(z) + \beta \cdot \lambda^{-1/2}\big[K(z,z) - k_{\mathcal{D}}(z)^\top (K_{\mathcal{D}} + \lambda I)^{-1} k_{\mathcal{D}}(z)\big]^{1/2}, H - h + 1\Big\}^+, \tag{4.3}$$

where $Q_0 \in \mathcal{H}$, similar to $\widehat{Q}_h^t$ in (3.4), is the solution to a kernel ridge regression problem and $\mathcal{D} \subseteq \mathcal{Z}$ is a discrete subset of $\mathcal{Z}$ with no more than T state-action pairs. Moreover, $K_{\mathcal{D}}$ and $k_{\mathcal{D}}$ are defined similarly as in (3.7) based on data in $\mathcal{D}$. Then, for any $R, B > 0$, we define a function class $\mathcal{Q}_{\mathrm{ucb}}(h, R, B)$ as

$$\mathcal{Q}_{\mathrm{ucb}}(h, R, B) = \big\{Q : Q \text{ takes the form of (4.3) with } \|Q_0\|_{\mathcal{H}} \le R, \beta \in [0, B], |\mathcal{D}| \le T\big\}. \tag{4.4}$$

As we will show in Lemma C.1, we have $\|\widehat{Q}_h^t\|_{\mathcal{H}} \le R_T$ for all $(t, h) \in [T] \times [H]$, where $R_T = 2H\sqrt{\Gamma_K(T, \lambda)}$. Thus, when B exceeds parameter β in (3.5), each Q_h^t is contained in $\mathcal{Q}_{\mathrm{ucb}}(h, R_T, B)$.

Moreover, since $r_h + \mathbb{P}_h V_{h+1}^t = \mathbb{T}_h^\star Q_{h+1}^t$ is the population ground truth of the ridge regression in (3.4), the complexity of $\mathcal{Q}_{\mathrm{ucb}}(h+1, R_T, B)$ naturally appears when quantifying the uncertainty of $\widehat{Q}_h^t$. To this end, for any $\epsilon > 0$, let $N_\infty(\epsilon; h, B)$ be the ϵ-covering number of $\mathcal{Q}_{\mathrm{ucb}}(h, R_T, B)$ with respect to the ℓ_∞-norm on $\mathcal{Z}$, which is also determined by the spectral structure of $\mathcal{H}$ and characterizes the complexity of the value functions constructed by KOVI.

Now we are ready to present the regret bound of KOVI.

Theorem 4.2. Assume that there exists $B_T > 0$ satisfying

$$8 \cdot \Gamma_K(T, 1 + 1/T) + 8 \cdot \log N_\infty(\epsilon^*; h, B_T) + 16 \cdot \log(2TH) + 22 + 2R_Q^2 \leq (B_T/H)^2 \tag{4.5}$$

for all $h \in [H]$, where $\epsilon^* = H/T$. We set $\lambda = 1 + 1/T$ and $\beta = B_T$ in Algorithm 2. Then, under Assumption 4.1, with probability at least $1 - (T^2H^2)^{-1}$, we have

$$\mathrm{Regret}(T) \leq 5\beta H \cdot \sqrt{T \cdot \Gamma_K(T, \lambda)}. \tag{4.6}$$

As shown in (4.16), the regret can be written as $\mathcal{O}(H^2 \cdot \delta_{\mathcal{H}} \cdot \sqrt{T})$, where $\delta_{\mathcal{H}} = B_T/H \cdot \sqrt{\Gamma_K(T, \lambda)}$ reflects the complexity of $\mathcal{H}$ and B_T satisfies (4.5). Specifically, $\delta_{\mathcal{H}}$ involves (i) the ℓ_∞-covering number $N_\infty(\epsilon^*, h, B_T)$ of $\mathcal{Q}_{\mathrm{ucb}}(h, R_T, B_T)$ and (ii) the effective dimension $\Gamma_K(T, \lambda)$, both characterize the intrinsic complexity of $\mathcal{H}$. Moreover, when neglecting the constant and logarithmic terms in (4.5), it suffices to choose B_T satisfying

$$B_T/H \asymp \sqrt{\Gamma_K(T, \lambda)} + \max_{h \in [H]} \sqrt{\log N_\infty(\epsilon^*, h, B_T)},$$

which reduces the regret bound in (4.16) to

$$\mathrm{Regret}(T) = \widetilde{\mathcal{O}}\Big(H^2 \cdot \underbrace{\Big[\Gamma_K(T, \lambda) + \max_{h \in [H]} \sqrt{\Gamma_K(T, \lambda) \cdot \log N_\infty(\epsilon^*, h, B_T)}\Big]}_{\delta_{\mathcal{H}}\text{: complexity of function class } \mathcal{H}} \cdot \sqrt{T}\Big). \tag{4.7}$$

To further obtain some intuition of (4.7), let us consider the tabular case where $\mathcal{Q}^\star$ consists of all measurable functions defined on $\mathcal{S} \times \mathcal{A}$ with range $[0, H]$. In this case, the value function class $\mathcal{Q}_{\mathrm{ucb}}(h, R_T, B_T)$ can be set to $\mathcal{Q}^\star$, whose ℓ_∞-covering number $N_\infty(\epsilon^*, h, B_T) \leq |\mathcal{S} \times \mathcal{A}| \cdot \log T$. Moreover, it can be shown that the effective dimension is also $\mathcal{O}(|\mathcal{S} \times \mathcal{A}| \cdot \log T)$. Thus, ignoring the logarithmic terms, Theorem 4.2 implies that by choosing $\beta \asymp H \cdot |\mathcal{S} \times \mathcal{A}|$, optimistic least-squares value iteration achieves an $\widetilde{\mathcal{O}}(H^2 \cdot |\mathcal{S} \times \mathcal{A}| \cdot \sqrt{T})$ regret.

Furthermore, we remark that the regret bound in (4.16) holds for any RKHS in general. It hinges on (i) Assumption 4.1, which postulates that the RKHS-norm ball $\{f \in \mathcal{H} \colon \|f\|_{\mathcal{H}} \leq R_Q H\}$ contains the image of the Bellman operator, and (ii) the inequality in (4.5) admits a solution B_T, which is set to be β in Algorithm 2. Here we set β to be sufficiently large so as to dominate the uncertainty of $\widehat{Q}_h^t$, whereas to quantify such uncertainty, we utilize the uniform concentration over the value function class $\mathcal{Q}_{\mathrm{ucb}}(h+1, R_T, \beta)$ whose complexity metric, the ℓ_∞-covering number, in turn depends on β. Such an intricate desideratum leads to (4.5) which determines β implicitly.

It is worth noting that the uniform concentration is unnecessary when $H = 1$. In this case, it suffices to choose $\beta = \widetilde{\mathcal{O}}(\sqrt{\Gamma_K(T, \lambda)})$ and KOVI incurs an $\widetilde{\mathcal{O}}(\Gamma_K(T, \lambda) \cdot \sqrt{T})$ regret, which matches the regret bounds of UCB algorithms for kernelized contextual bandits in Srinivas et al. (2009); Chowdhury and Gopalan (2017). Here $\widetilde{O}(\cdot)$ omits logarithmic terms. Thus, the covering number in (4.7) is specific for MDPs and arises due to the temporal dependence within an episode.

Furthermore, to obtain a concrete regret bound from (4.16), it remains to further characterize $\Gamma_K(T, \lambda)$ and $\log N_\infty(\epsilon^*, h, B_T)$ using characteristics of $\mathcal{H}$. To this end, in the following, we specify the eigenvalue decay property of $\mathcal{H}$.

Assumption 4.3 (Eigenvalue Decay of $\mathcal{H}$)**.** Recall that the integral operator T_K defined in (2.3) has eigenvalues $\{\sigma_j\}_{j\geq 1}$ and eigenfunctions $\{\psi_j\}_{j\geq 1}$. We assume that $\{\sigma_j\}_{j\geq 1}$ satisfies one of the following three eigenvalue decay conditions for some constant $\gamma > 0$:

(i) γ-finite spectrum: we have $\sigma_j = 0$ for all $j > \gamma$, where γ is a positive integer.

(ii) γ-exponential decay: there exist absolute constants C_1 and C_2 such that $\sigma_j \leq C_1 \cdot \exp(-C_2 \cdot j^\gamma)$ for all $j \geq 1$.

(iii) γ-polynomial decay: there exists a constant C_1 such that $\sigma_j \geq C_1 \cdot j^{-\gamma}$ for all $j \geq 1$, where $\gamma > 1$.

Moreover, for cases (ii) and (iii) above, we further assume that there exist constants $\tau \in [0, 1/2)$ $C_\psi > 0$ such that $\sup_{z\in\mathcal{Z}} \sigma_j^\tau \cdot |\psi_j(z)| \leq C_\psi$ for all $j \geq 1$.

Case (i) implies that $\mathcal{H}$ is a γ-dimensional RKHS. When this is the case, under Assumption 4.1, there exists a feature mapping $\phi\colon \mathcal{Z} \to \mathbb{R}^\gamma$ such that, for any $V\colon \mathcal{S} \to [0, H]$, $r_h + \mathbb{P}_h V$ is a linear function of ϕ. Such a property is satisfied by the linear MDP model studied in Yang and Wang (2019a,b); Jin et al. (2019); Zanette et al. (2020a). Moreover, when $\mathcal{H}$ satisfies case (i), KOVI reduces to the LSVI-UCB algorithm studied in Jin et al. (2019). In addition, cases (ii) and (iii) postulate that the eigenvalues of T_K decays exponentially and polynomially fast respectively, where γ is a constant that might depend on the input dimension d, which is assumed fixed throughout this paper. For example, the squared exponential kernel belongs to case (ii) with $\gamma = 1/d$, whereas the Matérn kernel with parameter $\nu > 2$ belongs to case (iii) with $\gamma = 1 + 2\nu/d$ (Srinivas et al., 2009). Moreover, we assume that there exists $\tau \in [0, 1/2)$ such that $\sigma_j^\tau \cdot \|\psi_j\|_\infty$ is universally bounded. Since $K(z, z) \leq 1$, this condition is naturally satisfied for $\tau = 1/2$. However, here we assume that $\tau \in (0, 1/2)$, which is satisfied when the magnitudes of the eigenvectors do grow not too fast compared with the decay of the eigenvalues. Such a condition is significantly weaker than assuming $\|\psi_j\|_\infty$ is universally bounded, which is also commonly made in the literature of non-parametric statistics (Lafferty and Lebanon, 2005; Shang et al., 2013; Zhang et al., 2015; Lu et al., 2016; Yang et al., 2017). It can be shown that the squared exponential kernel on unit sphere in $\mathbb{R}^d$ satisfy this condition for any $\tau > 0$ (See §B.3.) and the Matérn kernel on $[0, 1]$ satisfy this property with $\tau = 0$ (Yang et al., 2017). See Mendelson et al. (2010) for a more detailed discussion.

Now we present the regret bounds for the three eigenvalue decay conditions separately.

Corollary 4.4. Under Assumptions 4.1 and 4.3, we set $\lambda = 1 + 1/T$ and $\beta = B_T$ in Algorithm 2, where B_T is defined as

$$B_T = \begin{cases} C_b \cdot \gamma H \cdot \sqrt{\log(\gamma \cdot TH)} & \gamma\text{-finite spectrum}, \\ C_b \cdot H \sqrt{\log(TH)} \cdot (\log T)^{1/\gamma} & \gamma\text{-exponential decay}, \\ C_b \cdot H \log(TH) \cdot T^{\kappa^*} & \gamma\text{-polynomial decay}. \end{cases} \tag{4.8}$$

Here C_b is an absolute constant that does not depend on T or H, and κ^* is given by

$$\kappa^* = \max\left\{\xi^*, \frac{2d+\gamma+1}{(d+\gamma)\cdot[\gamma(1-2\tau)-1]}, \frac{2}{\gamma(1-2\tau)-3}\right\}, \qquad \xi^* = \frac{d+1}{2(\gamma+d)}. \tag{4.9}$$

For the third case, we further assume that γ is sufficiently large such that $\kappa^* + \xi^* < 1/2$. Then, there exists an absolute constant C_r such that, with probability at least $1 - (T^2H^2)^{-1}$, we have

$$\text{Regret}(T) \le \begin{cases} C_r \cdot H^2 \cdot \sqrt{\gamma^3 T} \cdot \log(\gamma TH) & \gamma\text{-finite spectrum}, \\ C_r \cdot H^2 \cdot \sqrt{(\log T)^{3/\gamma} \cdot T} \cdot \log(TH) & \gamma\text{-exponential decay}, \\ C_r \cdot H^2 \cdot T^{\kappa^*+\xi^*+1/2} \cdot \big[\log(TH)\big]^{3/2} & \gamma\text{-polynomial decay}. \end{cases} \tag{4.10}$$

Corollary 4.4 asserts that when β is chosen properly according to the eigenvalue decay property of $\mathcal{H}$, KOVI incurs a sublinear regret under all of the three cases specified in Assumption 4.3. Note that the linear MDP (Jin et al., 2019) satisfies the γ-finite spectrum condition and KOVI recovers the LSVI-UCB algorithm studied in Jin et al. (2019) when restricted to this setting. Moreover, our $\widetilde{\mathcal{O}}(H^2 \cdot \sqrt{\gamma^3 T})$ also matches the regret bound in Jin et al. (2019). In addition, under the γ-exponential eigenvalue decay condition, as we will show in §D, the log-covering number and the effective dimension are bounded by $(\log T)^{1+2/\gamma}$ and $(\log T)^{1+1/\gamma}$, respectively. Plugging these facts into (4.7), we obtain the sublinear regret in (4.16). As a concrete example, for the squared exponential kernel, we obtain an $\mathcal{O}(H^2 \cdot (\log T)^{1+1.5d} \cdot \sqrt{T})$ regret, where d is the input dimension. This such a regret is $(\log T)^{d/2}$ worse than that in Srinivas et al. (2009) for kernel contextual bandits, which is due to bounding the log-covering number. Furthermore, for the case of γ-polynomial decay, since the eigenvalues decay to zero rather slowly, we fail to obtain a $\sqrt{T}$-regret and only obtain a sublinear regret in (4.16). As we will show in the proof, the log-covering number and the effective dimension are $\widetilde{\mathcal{O}}(T^{2\kappa^*})$ and $\widetilde{\mathcal{O}}(T^{2\xi^*})$, respectively, which, combined with (4.7), yields the regret bound in (4.16). As a concrete example, consider the Matérn kernel with parameter $\nu > 0$ where we have $\gamma = 1+2\nu/d$ and $\tau = 0$. In this case, when ν is sufficiently large such that $T^{2\xi^*-1/2} = o(1)$, we have

$$\xi^* = \frac{d(d+1)}{2[2\nu + d(d+1)]}, \qquad \kappa^* = \max\left\{\xi^*, \frac{3}{d-1}, \frac{2}{d-1}\right\},$$

which implies that KOVI achieves an $\widetilde{\mathcal{O}}(H^2 \cdot T^{2\xi^*+1/2})$ regret when d is large. Such a regret bound matches that in Srinivas et al. (2009) for the bandit setting. See §B.1 for details.

Furthermore, similarly to the discussion in Section 3.1 of Jin et al. (2018), the regret bound in (4.16) directly translates to an upper bound on the sample complexity as follows. When the initial state is fixed for all episodes, for any fixed $\epsilon > 0$, with at least a constant probability, KOVI returns a policy π satisfying $V_1^\star(x_1) - V_1^\pi(x_1) \le \epsilon$ using $\mathcal{O}(H^4 B_T^2 \cdot \Gamma_K(T,\lambda)/\epsilon^2)$ samples. Specifically, for the three cases considered in Assumption 4.3, such a sample complexity guarantee reduces to

$$\widetilde{\mathcal{O}}\big(H^4 \cdot \gamma^3/\epsilon^2\big), \qquad \widetilde{\mathcal{O}}\big(H^4 \cdot (\log T)^{2+3/\gamma}/\epsilon^2\big), \qquad \widetilde{\mathcal{O}}\big(H^4 \cdot T^{2(\kappa^*+\xi^*)}/\epsilon^2\big),$$

respectively. Moreover, similar to Jin et al. (2019), our analysis can also be extended to the misspecified setting where $\inf_{f\in\mathcal{Q}^\star}\|f - \mathcal{T}_h^\star Q\|_\infty \le \mathtt{err}_{\mathrm{mis}}$ for all $Q\colon \mathcal{Z} \to [0,H]$. Here $\mathtt{err}_{\mathrm{mis}}$ is the model misspecification error. Under this setting, KOVI will suffer from an extra $\mathtt{err}_{\mathrm{mis}} \cdot TH$ regret. The analysis for the misspecified setting is similar to that for the neural setting that will be presented in §4.2. Hence we omit the analysis details for brevity.

4.2 Regret of NOVI

In this section, we establish the regret of NOVI. Throughout this subsection, we let $\mathcal{H}$ be the RKHS whose kernel function is K_{ntk} define in (2.9). Also recall that we regard $\mathcal{Z} = \mathcal{S}\times\mathcal{A}$ as a subset of the unit sphere $\mathbb{S}^{d-1} = \{z \in \mathbb{R}^d\colon \|z\|_2 = 1\}$. Moreover, let $(b^{(0)}, W^{(0)})$ be the initial value of the network weights obtained via the symmetric initialization scheme introduced in §2.3. Conditioning on the randomness of the initialization, we define a finite-rank kernel $K_m\colon \mathcal{Z}\times\mathcal{Z}\to\mathbb{R}$ by letting $K_m(z,z') = \langle \nabla_W f(z; b^{(0)}, W^{(0)}), \nabla_W f(z'; b^{(0)}, W^{(0)})\rangle$. Notice that the rank of K_m is md, where m is much larger than T and H and is allowed to increase to infinity. Besides, with a slight abuse of notation, we define

$$\mathcal{Q}^\star = \bigg\{ f_\alpha(z) = \int_{\mathbb{R}^d} \mathrm{act}'(w^\top z)\cdot z^\top \alpha(w)\ \mathrm{d}p_0(w)\colon \alpha\colon \mathbb{R}^d \to \mathbb{R}^d, \|\alpha\|_{2,\infty} \le R_Q H/\sqrt{d} \bigg\}, \tag{4.11}$$

where R_Q is a positive number, p_0 is the density of $N(0, I_d/d)$, and we define $\|\alpha\|_{2,\infty} = \sup_w \|\alpha(w)\|_2$. That is, $\mathcal{Q}^\star$ consists of functions that can be expressed as infinite number of random features. As shown in Lemma C.1 of Gao et al. (2019), $\mathcal{Q}^\star$ is a dense subset of the RKHS $\mathcal{H}$. Thus, when R_Q is sufficiently large, $\mathcal{Q}^\star$ in (4.11) is an expressive function class. We impose the following condition on $\mathcal{Q}^\star$.

Assumption 4.5. We assume that for any $h\in[H]$ and any $Q\colon \mathcal{S}\times\mathcal{A}\to[0,H]$, we have $\mathbb{T}_h^\star Q \in \mathcal{Q}^\star$.

Assumption 4.5 is in the same vein as Assumption 4.1. Here we focus on $\mathcal{Q}^*$ instead of an RKHS norm ball of NTK only due to technical considerations. However, since functions of the form in (4.11) are dense in $\mathcal{H}$, Assumptions 4.5 and 4.1 are indeed very similar.

To characterize the value function class associated with NOVI, for any discrete set $\mathcal{D}\subseteq\mathcal{Z}$, similar to (3.11), we define

$$\overline{\Lambda}_{\mathcal{D}} = \lambda\cdot I_{2md} + \sum_{z\in\mathcal{D}} \varphi(z; W^{(0)})\varphi(z; W^{(0)})^\top,$$

where $\varphi(\cdot; W^{(0)})$ is the neural tangent feature defined in (2.6). With a slight abuse of notation, for any $R, B > 0$, we let $\mathcal{Q}_{\mathrm{ucb}}(h, R, B)$ denote that class of functions that take the form of

$$Q(z) = \min\Big\{ \big\langle \varphi(z; W^{(0)}), W\big\rangle + \beta\cdot\big[\varphi(z;W^{(0)})^\top(\overline{\Lambda}_{\mathcal{D}})^{-1}\varphi(z;W^{(0)})\big]^{1/2}, H-h+1\Big\}^+, \tag{4.12}$$

where $W \in \mathbb{R}^{2md}$ satisfies $\|W\|_2 \le R$, $\beta \in [0, B]$, and $\mathcal{D}$ has cardinality no more than T. Intuitively, when both R and B are sufficiently large, $\mathcal{Q}_{\mathrm{ucb}}(h, R, B)$ contains the counterpart of neural-based value function Q_h^t that is based on neural tangent features. When m is sufficiently large, it is expected that Q_h^t is well-approximately by functions in $\mathcal{Q}_{\mathrm{ucb}}(h, R, B)$ where the approximation error decays with m. It is worth noting the class of linear functions of $\varphi(\cdot; W^{(0)})$ forms an RKHS with kernel K_m in (2.7). Any function f in this class can be written as $f(\cdot) = \langle \varphi(\cdot; W^{(0)}), W_f \rangle$ for some $W_f \in \mathbb{R}^{2md}$. Moreover, the RKHS norm of f is given by $\|W_f\|_2$. Thus, $\mathcal{Q}_{\mathrm{ucb}}(h, R, B)$ defined above coincides with the counterpart defined in (4.4) with the kernel function being K_m. We set $R_T = H\sqrt{2T/\lambda}$ and let $N_\infty(\epsilon; h, B)$ denote the ϵ-covering number of $\mathcal{Q}_{\mathrm{ucb}}(h, R_T, B)$ with respect to the ℓ_∞-norm on $\mathcal{Z}$.

In the following theorem, we present a general regret bound for NOVI.

Theorem 4.6. Under Assumptions 4.5, We also assume that m is sufficiently large such that $m = \Omega(T^{13}H^{14} \cdot (\log m)^3)$. In Algorithm 3, we let λ be a sufficiently large constant and let $\beta = B_T$ which satisfies inequality

$$16\Gamma_{K_m}(T, \lambda) + 16 \cdot \log N_\infty(\epsilon^*, h+1, B_T) + 32 \cdot \log(2TH) + 4R_Q^2 \cdot (1 + \lambda/d) \le (B_T/H)^2 \quad (4.13)$$

for all $h \in [H]$. Here $\epsilon^* = H/T$ and $\Gamma_{K_m}(T, \lambda)$ is the maximal information gain defined for kernel K_m. In addition, for the neural network in (2.5), we assume the activation function act is C_{act}-smooth, i.e., its derivative act$'$ is C_{act}-Lipschitz, and m is sufficiently large such that

$$m = \Omega\big(\beta^{12} \cdot T^{13} \cdot H^{14} \cdot (\log m)^3\big). \quad (4.14)$$

Then with probability at least $1 - (T^2H^2)^{-1}$, we have

$$\mathrm{Regret}(T) = 5\beta H \cdot \sqrt{T \cdot \Gamma_{K_m}(T, \lambda)} + 10\beta TH \cdot \iota, \quad (4.15)$$

where we define $\iota = T^{7/12} \cdot H^{1/6} \cdot m^{-1/12} \cdot (\log m)^{1/4}$.

This theorem shows that, when m is sufficiently large, NOVI enjoys a similar regret bound as KOVI. Specifically, the choice of β in (4.13) is similar to that in (4.5) for kernel K_m. Here we set λ to be an absolute constant as $\sup_z K_m(z, z) \le 1$ no longer holds. In addition, here we assume that act$'$ is C_{act}-Lipschitz on $\mathbb{R}$, which can be relaxed to only assuming act$'$ is Lipschitz continous on a bounded interval of $\mathbb{R}$ that contains $w^\top z$ with high probability, where w is drawn from the initial distribution of W_j, $j \in [m]$.

Moreover, comparing (4.16) and (4.15) we observe that, when m is sufficiently large, NOVI can be viewed as a misspecified version of KOVI for the RKHS with kernel K_m, where the model misspecification error is $\mathtt{err}_{\mathrm{mis}} = 10\beta \cdot \iota$. Specifically, the first term in (4.15) is the same as that in (4.16), where the choice of β and $\Gamma_{K_m}(T, \lambda)$ reflect the intrinsic complexity of K_m. Whereas the second term is equal to $\mathtt{err}_{\mathrm{mis}} \cdot TH$, which arises due to approximating neural network value functions by functions in $\mathcal{Q}_{\mathrm{ucb}}(h, R_T, B_T)$, which are constructed using kernel functions with feature mapping $\varphi(\cdot; W^{(0)})$. Moreover, when β is bounded by a polynomial of TH, to make $\mathtt{err}_{\mathrm{mis}} \cdot TH$ negligible, it suffices to let m be a polynomial of TH. That is, when the network width is a polynomial of the total number of steps, NOVI achieves the same performance as KOVI.

Furthermore, when neglecting the constants and logarithmic terms in (4.13), we simplify the regret bound in (4.15) into

$$\text{Regret}(T) = \widetilde{\mathcal{O}}\Big(H^2 \cdot \Big[\Gamma_{K_m}(T,\lambda) + \max_{h\in[H]} \sqrt{\Gamma_{K_m}(T,\lambda) \cdot \log N_\infty(\epsilon^*, h, B_T)}\Big] \cdot \sqrt{T} + \texttt{err}_{\text{mis}} \cdot T\Big).$$

which depends on the intrinsic complexity of K_m through both the effective dimension $\Gamma_{K_m}(T,\lambda)$ and the log-covering number $\log N_\infty(\epsilon^*, h, B_T)$. To obtain a more concrete regret bounds, in the following, we pose an assumption on the spectral structure of K_m.

Assumption 4.7 (Eigenvalue Decay of the Empirical NTK)**.** Conditioning on the randomness of $(b^{(0)}, W^{(0)})$, let K_m be the kernel induced by the neural tangent features $\nabla f(\cdot; b^{(0)}, W^{(0)})$. Let T_{K_m} be the integral operator induced by K_m and the Lebesgue measure on $\mathcal{Z}$ and let $\{\sigma_j\}_{j\geq 1}$ and $\{\psi_j\}_{j\geq 1}$ be its eigenvalues and eigenvectors, respectively. We assume that $\{\sigma_j\}_{j\geq 1}$ and $\{\psi_j\}_{j\geq 1}$ satisfy either one of the three decay conditions specified in Assumption 4.3. Here we assume the constants C_1, C_2, C_ψ, γ, and τ do not depend on m.

Here we assume that K_m satisfies Assumption 4.3. Since K_m depends on the initial network weights, which are random, this assumption should be better understood in the limit sense. Specifically, as m goes to infinity, K_m converges to K_{ntk}, which is determined by both the activation function and the distribution of the initial network weights. Thus, if the RKHS with kernel K_{ntk} satisfy Assumption 4.3, when m is sufficiently large, it is reasonable to expect that such a condition also holds for K_m. Due to the space limit, we present concrete examples of K_{ntk} satisfying Assumption 4.3 in §B.3 in the appendix.

Now we are ready to characterize the performances of NOVI for each case separately.

Corollary 4.8. Under Assumptions 4.5 and 4.7, we assume the activation function is C_{act}-smooth and the number of neurons of the neural network satisfies (4.14). Besides, in Algorithm 3 we let λ be a sufficiently large constant and set $\beta = B_T$ as in (4.8). Then exists an absolute constant C_r such that, with probability at least $1 - (T^2H^2)^{-1}$, we have

$$\text{Regret}(T) \leq \begin{cases} C_r \cdot H^2 \cdot \sqrt{\gamma^3 T} \cdot \log(\gamma TH) + 10\beta TH \cdot \iota & \gamma\text{-finite spectrum}, \\ C_r \cdot H^2 \cdot \sqrt{(\log T)^{3/\gamma} \cdot T} \cdot \log(TH) + 10\beta TH \cdot \iota & \gamma\text{-exponential decay}, \\ C_r \cdot H^2 \cdot T^{\kappa^*+\xi^*+1/2} \cdot \big[\log(TH)\big]^{3/2} + 10\beta TH \cdot \iota & \gamma\text{-polynomial decay}, \end{cases} \tag{4.16}$$

where we define $\iota = T^{7/12} \cdot H^{1/6} \cdot m^{-1/12} \cdot (\log m)^{1/4}$.

Corollary 4.8 is parallel to Corollary 4.4, with an additional misspecification error $10\beta TH \cdot \iota$. It remains to see whether there exist concrete neural networks that induce NTKs satisfying each eigenvalue decay condition. As we will show in §B.3, neural network with quadratic activation function induces an NTK with a finite spectrum, while the sine activation function and the polynomials of ReLU activations induce NTKs satisfying the exponential and polynomial eigenvalue decay conditions, respectively. Corollary 4.8 can be directly applied to these concrete examples to obtain sublinear regret bounds.

5 Proofs of the Main Results

In this section, we provide the proofs of Theorems 4.2 and 4.6. The proofs of the supporting lemmas and auxiliary results are deferred to the appendix.

5.1 Proof of Theorem 4.2

Proof. For the simplicity of presentation, we define the temporal-difference (TD) error as

$$\delta_h^t(x,a) = r_h(x,a) + (\mathbb{P}_h V_{h+1}^t)(x,a) - Q_h^t(x,a), \qquad \forall (x,a) \in \mathcal{S} \times \mathcal{A}. \tag{5.1}$$

Here δ_h^t is a function on $\mathcal{S} \times \mathcal{A}$ for all $h \in [H]$ and $t \in [T]$. Note that $V_h^t(\cdot) = \max_{a \in \mathcal{A}} Q_h^t(\cdot, a)$. Intuitively, $\{\delta_h^t\}_{h\in[H]}$ quantifies the how far $\{Q_h^t\}_{h\in[H]}$ are from satisfying the Bellman optimality equation in (2.2). In addition, recall that π^t is the policy executed in the t-th episode, which generates a trajectory $\{(x_h^t, a_h^t)\}_{h\in[H]}$. For any $h \in [H]$ and $t \in [T]$, we further define $\zeta_{t,h}^1, \zeta_{t,h}^2 \in \mathbb{R}$ as

$$\zeta_{t,h}^1 = \big[V_h^t(x_h^t) - V_h^{\pi^t}(x_h^t)\big] - \big[Q_h^t(x_h^t, a_h^t) - Q_h^{\pi^t}(x_h^t, a_h^t)\big], \tag{5.2}$$

$$\zeta_{t,h}^2 = \big[(\mathbb{P}_h V_{h+1}^t)(x_h^t, a_h^t) - (\mathbb{P}_h V_{h+1}^{\pi^t})(x_h^t, a_h^t)\big] - \big[V_{h+1}^t(x_{h+1}^t) - V_{h+1}^{\pi^t}(x_{h+1}^t)\big]. \tag{5.3}$$

By definition, $\zeta_{t,h}^1$ and $\zeta_{t,h}^2$ captures two sources of randomness, respectively — the randomness of choosing an action $a_h^t \sim \pi_h^t(\cdot \,|\, x_h^t)$ and that of drawing the next state x_{h+1}^t from $\mathbb{P}_h(\cdot \,|\, x_h^t, a_h^t)$. As we will see in §C.3, $\{\zeta_{t,h}^1, \zeta_{t,h}^2\}$ form a bounded Martingale difference sequence with respect to a properly chosen filtration, which enables us to bound their total sum via the Azuma-Hoeffding inequality (Azuma, 1967).

To establish an upper bound of the regret, in the following lemma, we first decompose the regret into three parts using the notations defined above. Similar regret decomposition results also appear in Cai et al. (2019a); Efroni et al. (2020).

Lemma 5.1 (Regret Decomposition)**.** Notice that we the temporal-difference error $\delta_h^t \colon \mathcal{S} \times \mathcal{A} \to$ in (5.1) for all $(t,h) \in [T] \times [H]$. Then we can write the regret as

$$\mathrm{Regret}(T) = \underbrace{\sum_{t=1}^T \sum_{h=1}^H \big[\mathbb{E}_{\pi^\star}[\delta_h^t(x_h, a_h) \,|\, x_1 = x_1^t] - \delta_h^t(x_h^t, a_h^t)\big]}_{\text{(i)}} + \underbrace{\sum_{t=1}^T \sum_{h=1}^H (\zeta_{t,h}^1 + \zeta_{t,h}^2)}_{\text{(ii)}}$$
$$\underbrace{\sum_{t=1}^T \sum_{h=1}^H \mathbb{E}_{\pi^\star}\big[\big\langle Q_h^t(x_h, \cdot), \pi_h^\star(\cdot \,|\, x_h) - \pi_h^t(\cdot \,|\, x_h)\big\rangle_{\mathcal{A}} \,\big|\, x_1 = x_1^t\big]}_{\text{(iii)}}, \tag{5.4}$$

where $\zeta_{t,h}^1$ and $\zeta_{t,h}^2$ are defined in (5.2) and (5.3), respectively.

Proof. See §C.1 for a detailed proof. ☐

To begin with, notice that π_h^t is the greedy policy with respect to Q_h^t for all $(t,h) \in [T] \times [H]$. By definition, we have

$$\big\langle Q_h^t(x_h,\cdot), \pi_h^\star(\cdot \,|\, x_h) - \pi_h^t(\cdot \,|\, x_h)\big\rangle_{\mathcal{A}} = \big\langle Q_h^t(x_h,\cdot), \pi_h^\star(\cdot \,|\, x_h)\big\rangle_{\mathcal{A}} - \max_{a\in\mathcal{A}} Q_h^t(x_h,a) \le 0$$

for all $x_h \in \mathcal{S}$. Thus, Term (iii) in (5.4) is non-positive. Then, by Lemma 5.1, we can upper bound the regret by

$$\text{Regret}(T) \le \underbrace{\bigg\{\sum_{t=1}^T \sum_{h=1}^H \big[\mathbb{E}_{\pi^\star}[\delta_h^t(x_h,a_h) \,|\, x_1 = x_1^t] - \delta_h^t(x_h^t,a_h^t)\big]\bigg\}}_{\text{(i)}} + \underbrace{\bigg[\sum_{t=1}^T \sum_{h=1}^H (\zeta_{t,h}^1 + \zeta_{t,h}^2)\bigg]}_{\text{(ii)}}. \tag{5.5}$$

For Term (i), since we do not observe trajectories from π^*, which is unknown, it appears that $\mathbb{E}_{\pi^*}[\delta_h^t(x_h,a_h) \,|\, x_1 = x_1^t]$ cannot be estimated. Fortunately, by adding a bonus term in Algorithm 2, we ensure that the temporal-difference error δ_h^t is always a non-positive function, which is shown in the following lemma.

Lemma 5.2 (Optimism)**.** Let $\lambda = 1 + 1/T$ and $\beta = B_T$ in Algorithm 2, where B_T satisfies (4.5). Under Assumptions 4.1, with probability at least $1 - (2T^2H^2)^{-1}$, it holds for all $(t,h) \in [T] \times [H]$ and $(x,a) \in \mathcal{S} \times \mathcal{A}$ that

$$-2\beta \cdot b_h^t(x,a) \le \delta_h^t(x,a) \le 0.$$

Proof. See §C.2 for a detailed proof. □

Applying Lemma 5.2 to Term (i) in (5.5), we obtain that

$$\text{Term (i)} \le \bigg[\sum_{t=1}^T \sum_{h=1}^H -\delta_h^t(x_h^t,a_h^t)\bigg] \le 2\beta \cdot \bigg[\sum_{t=1}^T \sum_{h=1}^H b_h^t(x_h^t,a_h^t)\bigg] \tag{5.6}$$

holds with probability at least $1 - (2T^2H^2)^{-1}$, where β is equal to B_T that specified in (4.5).

Finally, it remains to bound the sum of bonus terms in (5.6). As we show in (C.17), using the feature representation of $\mathcal{H}$, we can write each $b_h^t(x_h^t,a_h^t)$ as

$$b_h^t(x_h^t,a_h^t) = \big[\phi(x_h^t,a_h^t)^\top (\Lambda_h^t)^{-1} \phi(x_h^t,a_h^t)\big]^{1/2},$$

where $\Lambda_h^t = \lambda \cdot I_{\mathcal{H}} + \sum_{\tau=1}^{t-1} \phi(x_h^t,a_h^t)\phi(x_h^t,a_h^t)^\top$ is a self-adjoint and positive definite operator on $\mathcal{H}$ and $\mathcal{I}_{\mathcal{H}}$ is the identity mapping on $\mathcal{H}$. Thus, combining Cauchy-Schwarz inequality and Lemma E.3, for any $h \in [H]$, with probability at least $1 - (2T^2H^2)^{-1}$ we have

$$\begin{aligned} \text{Term (i)} &\le 2\beta \cdot \sqrt{T} \cdot \sum_{h=1}^H \bigg[\sum_{t=1}^T \phi(x_h^t,a_h^t)^\top (\Lambda_h^t)^{-1} \phi(x_h^t,a_h^t)\bigg]^{1/2} \\ &\le 2\beta \cdot \sum_{h=1}^H \big[2T \cdot \operatorname{logdet}(I + K_h^T/\lambda)\big]^{1/2} = 4\beta H \cdot \sqrt{T \cdot \Gamma_K(T,\lambda)}, \end{aligned} \tag{5.7}$$

where $\Gamma_K(T,\lambda)$ is the maximal information gain defined in (4.2) with parameter λ.

It remains to bound Term (ii) in (5.5), which is given by the following lemma.

Lemma 5.3. For $\zeta^1_{t,h}$ and $\zeta^2_{t,h}$ defined respectively in (5.2) and (5.3) and for any $\zeta \in (0,1)$, with probability at least $1-\zeta$, we have

$$\sum_{t=1}^{T}\sum_{h=1}^{H}(\zeta^1_{t,h}+\zeta^2_{t,h}) \le \sqrt{16TH^3\cdot\log(2/\zeta)}.$$

Proof. See §C.3 for a detailed proof. ☐

Setting $\zeta = (2T^2H^2)^{-1}$ in Lemma 5.3 we obtain that

$$\text{Term (ii)} = \sum_{t=1}^{T}\sum_{h=1}^{H}(\zeta^1_{t,h}+\zeta^2_{t,h}) \le \sqrt{16TH^3\cdot\log(4T^2H^2)} = \sqrt{32TH^3\cdot\log(2TH)} \tag{5.8}$$

holds with probability at least $1-(2TH)^{-1}$.

Therefore, combining (4.5), (5.5), and (5.8), we conclude that, with probability at least $1-(T^2H^2)^{-1}$, the regret is bounded by

$$\text{Regret}(T) \le 4\beta H\cdot\sqrt{T\cdot\Gamma_K(T,\lambda)} + \sqrt{32TH^3\cdot\log(2TH)} \le 5\beta H\cdot\sqrt{T\cdot\Gamma_K(T,\lambda)},$$

where the last inequality follows from the choice of $\beta = B_T$, which implies that

$$\beta \ge H\cdot\sqrt{16\log(TH)} \ge \sqrt{32H\cdot\log(2TH)}.$$

Therefore, we conclude the proof of Theorem 4.2. ☐

5.2 Proof of Theorem 4.6

Proof. The proof of Theorem 4.6 is similar to that of Theorem 4.2. Recall that we let $\mathcal{Z}$ denote $\mathcal{S}\times\mathcal{A}$ for simplicity. Moreover, also recall that for all $(t,h)\in[T]\times[H]$, we define the temporal-difference (TD) error $\delta^t_h\colon \mathcal{Z}\to\mathbb{R}$ in (5.1) and define random variables $\zeta^1_{t,h}$ and $\zeta^2_{t,h}$ in (5.2) and (5.3), respectively.

Then, combining Lemma 5.1 and the fact that π^t is the greedy policy with respect to $\{Q^t_h\}_{h\in[H]}$, we bound the regret by

$$\text{Regret}(T) \le \underbrace{\left\{\sum_{t=1}^{T}\sum_{h=1}^{H}\left[\mathbb{E}_{\pi^\star}[\delta^t_h(x_h,a_h)\,|\,x_1=x^t_1] - \delta^t_h(x^t_h,a^t_h)\right]\right\}}_{\text{(i)}} + \underbrace{\left[\sum_{t=1}^{T}\sum_{h=1}^{H}(\zeta^1_{t,h}+\zeta^2_{t,h})\right]}_{\text{(ii)}}. \tag{5.9}$$

Here, Term (ii) is a sum of Martingale difference sequence. By setting $\zeta = (4T^2H^2)^{-1}$ in Lemma 5.3, with probability at least $1-(4T^2H^2)^{-1}$, we have

$$\text{Term (ii)} = \sum_{t=1}^{T}\sum_{h=1}^{H}(\zeta^1_{t,h}+\zeta^2_{t,h}) \le \sqrt{16TH^3\cdot\log(8T^2H^2)} \le H\cdot\sqrt{32TH\log(2TH)}. \tag{5.10}$$

It remains to bound Term (i) in (5.9). To this end, we aim to establish a counterpart of Lemma 5.2 for neural value functions, which shows that, by adding a bonus term $\beta\cdot b^t_h$, the TD error δ^t_h

is always a non-positive function approximately. Then bounding Term (i) in (5.9) is reduced to handling $\sum_{t=1}^{T}\sum_{h=1}^{H} b_h^t(x_h^t, a_h^t)$.

Note that the bonus function b_h^t are constructed based on neural tangent feature $\varphi(\cdot; \widehat{W}_h^t)$ and matrix Λ_h^t. In order to relate $\sum_{t=1}^{T}\sum_{h=1}^{H} b_h^t(x_h^t, a_h^t)$ to the maximal information gain of the empirical NTK K_m, similar to Λ_h^t and b_h^t, we define $\overline{\Lambda}_h^t$ and $\overline{b}_h^t$ respectively as

$$\overline{\Lambda}_h^t = \lambda \cdot I_{2md} + \sum_{\tau=1}^{t-1} \varphi(x_h^\tau, a_h^\tau; W^{(0)})\varphi(x_h^\tau, a_h^\tau; W^{(0)})^\top, \qquad \overline{b}_h^t(z) = \big[\varphi(z; W^{(0)})^\top (\overline{\Lambda}_h^t)^{-1}\varphi(z; W^{(0)})\big]^{1/2}.$$

In the following lemma, we bound the TD-error δ_h^t using $\overline{b}_h^t$ and show that b_h^t and $\overline{b}_h^t$ are close in the ℓ_∞-norm on $\mathcal{Z}$ when m is sufficiently large.

Lemma 5.4 (Optimism)**.** Let λ be an absolute constant and $\beta = B_T$ in Algorithm 3, where B_T satisfies (4.13). Under the assumptions made in Theorem 4.6, with probability at least $1 - (2T^2H^2)^{-1} - m^2$, it holds for all $(t, h) \in [T] \times [H]$ and $(x, a) \in \mathcal{S} \times \mathcal{A}$ that

$$-5\beta \cdot \iota - 2\beta \cdot \overline{b}_h^t(x, a) \le \delta_h^t(x, a) \le 5\beta \cdot \iota, \qquad \sup_{(x,a)\in\mathcal{Z}} \big|b_h^t(x, a) - \overline{b}_h^t(x, a)\big| \le 2\iota, \tag{5.11}$$

where we define $\iota = T^{7/12} \cdot H^{1/12} \cdot m^{-1/12} \cdot (\log m)^{1/4}$.

Proof. See §C.4 for a detailed proof. □

Applying Lemma 5.2 to Term (i) in (5.5), we obtain that

$$\text{Term (i)} \le \Bigg[\sum_{t=1}^{T}\sum_{h=1}^{H} -\delta_h^t(x_h^t, a_h^t)\Bigg] + 5TH \cdot \iota \le 2\beta \cdot \Bigg[\sum_{t=1}^{T}\sum_{h=1}^{H} \overline{b}_h^t(x_h^t, a_h^t)\Bigg] + 10\beta TH \cdot \iota \tag{5.12}$$

holds with probability at least $1 - (2T^2H^2)^{-1} - m^{-2}$, where $\beta = B_T$. Moreover, combining Cauchy-Schwarz inequality and Lemma E.3, we have

$$\begin{aligned}\sum_{t=1}^{T}\sum_{h=1}^{H} \overline{b}_h^t(x_h^t, a_h^t) &\le \sqrt{T} \cdot \sum_{h=1}^{H} \Bigg[\sum_{t=1}^{T} \varphi(x_h^t, a_h^t; W^{(0)})^\top (\overline{\Lambda}_h^t)^{-1} \varphi(x_h^t, a_h^t; W^{(0)})\Bigg]^{1/2} \\ &\le 2H \cdot \sqrt{T \cdot \Gamma_{K_m}(T, \lambda)},\end{aligned} \tag{5.13}$$

where $\Gamma_K(T, \lambda)$ is the maximal information gain defined in (4.2) for kernel K_m.

Notice that $(2T^2H^2)^{-1} + m^{-2} + (4T^2H^2)^{-1} \le (T^2H^2)^{-1}$. Thus, combining (5.9), (5.10), (5.12), and (5.13), we obtain that

$$\begin{aligned}\text{Regret}(T) &\le 4\beta H \cdot \sqrt{T \cdot \Gamma_{K_m}(T, \lambda)} + 10\beta TH \cdot \iota + H \cdot \sqrt{32TH \log(2TH)} \\ &\le 5\beta H \cdot \sqrt{T \cdot \Gamma_{K_m}(T, \lambda)} + 10\beta TH \cdot \iota\end{aligned}$$

holds with probability at least $1 - (2T^2H^2)^{-1}$. Here the last inequality follows from the fact that

$$\beta \ge H \cdot \sqrt{32 \log(TH)} \ge \sqrt{32H \log(2TH)}.$$

Therefore, we conclude the proof of Theorem 4.6. □

6 Conclusion

In this paper, we have presented the algorithmic framework of optimistic least-squares value iteration for RL with general function approximation, where we propose to add an additional bonus term to the solution to each least-squares value estimation problem to promote exploration. Moreover, when applying this framework to the settings with kernel functions and overparameterized neural networks, we obtain two algorithms, KOVI and NOVI, respectively, that both provably enjoy sample and computational efficiency. Specifically, under the kernel and neural settings respectively, KOVI and NOVI both achieve sublinear $\widetilde{O}(\delta_{\mathcal{F}} H^2 \sqrt{T})$ regret upper bounds, where $\delta_{\mathcal{F}}$ is a quantity that characterizes the intrinsic complexity of the function class $\mathcal{F}$. To the best of our knowledge, we have developed the first provably efficient RL algorithms under the settings of kernel and neural function approximations.

Acknowledgements

Mengdi Wang gratefully acknowledges funding from the U.S. National Science Foundation (NSF) grant CMMI1653435, Air Force Office of Scientific Research (AFOSR) grant FA9550-19-1-020, and C3.ai DTI. Zhuoran Yang would like to thank Jianqing Fan, Csaba Szepsvári, Tuo Zhao, Simon Shaolei Du, Ruosong Wang, and Yiping Lu for valuable discussions.

A Neural Optimistic Least-Squares Value Iteration

In this section, we lay out the details of NOVI, which is omitted for brevity. We remark that the loss function L_h^t in Line 7 is given in (3.9) and its global minimizer $\widehat{W}_h^t$ can be efficiently obtained by first-order optimization methods.

Algorithm 3 Neural Optimistic Least-Squares Value Iteration (NOVI)

1: **Input:** Parameters λ and β.
2: Initialize the network weights $(b^{(0)}, W^{(0)})$ via the symmetric initialization scheme.
3: **for** episode $t = 1, \ldots, T$ **do**
4:     Receive the initial state x_1^t.
5:     Set V_{H+1}^t as the zero function.
6:     **for** step $h = H, \ldots, 1$ **do**
7:         Solve the neural network optimization problem $\widehat{W}_h^t = \mathrm{argmin}_W\, L_h^t(W)$.
8:         Update $\Lambda_h^t = \Lambda_h^{t-1} + \varphi(x_h^{t-1}, a_h^{t-1}; \widehat{W}_h^t)\varphi(x_h^{t-1}, a_h^{t-1}; \widehat{W}_h^t)^\top$.
9:         Obtain the bonus function b_h^t defined in (3.12).
10:         Obtain value functions

$$Q_h^t(\cdot,\cdot) \leftarrow \min\bigl\{f\bigl(\cdot,\cdot;\widehat{W}_h^t\bigr) + \beta\cdot b_h^t(\cdot,\cdot), H-h+1\bigr\}^+, \qquad V_h^t(\cdot) = \max_a Q_h^t(\cdot, a).$$

11:     **end for**
12:     **for** step $h = 1, \ldots, H$ **do**
13:         Take action $a_h^t \leftarrow \mathrm{argmax}_{a\in\mathcal{A}}\, Q_h^t(x_h^t, a)$.
14:         Observe the reward $r_h(x_h^t, a_h^t)$ and the next state x_{h+1}^t.
15:     **end for**
16: **end for**

B Proofs of the Corollaries

In this section, we prove Corollaries 4.4 and 4.8, which establish the regret for KOVI and NOVI under each specific eigenvalue decay condition. Moreover, in §B.3 we provide concrete examples of neural tangent kernels that satisfy Assumption 4.3 and further apply Corollaries 4.4 and 4.8 to these examples.

B.1 Proof of Corollary 4.4

Proof. To prove this corollary, it suffices to verify that for each eigenvalue decay condition specified in Assumption 4.3, B_T defined in (4.8) satisfies the condition in (4.5). Recall that we set $\lambda = 1+1/T$ in Algorithm 2 and denote $R_T = 2H\sqrt{\Gamma_K(T,\lambda)}$, $\epsilon^* = H/T$. Also recall that we let $N_\infty(\epsilon, h, B)$ denote the ϵ-covering number of $\mathcal{Q}_{\mathrm{ucb}}(h, R_T, B)$ with respect to the ℓ_∞-norm. In the sequel, we consider the three cases separately.

Case (i): γ-Finite Spectrum. When $\mathcal{H}$ has at most γ nonzero eigenvalues, by Lemma D.5, we have $\Gamma_K(T,\lambda) \le C_K \cdot \gamma \log T$, where C_K is an absolute constant. Moreover, by Lemma D.1, for any $h \in [H]$, we have

$$\begin{aligned}\log N_\infty(\epsilon^*, h, B_T) &\le C_N \cdot \gamma \cdot \bigl\{1 + \log\bigl[2\sqrt{\Gamma(T,\lambda)} \cdot T\bigr]\bigr\} + C_N \cdot \gamma^2 \cdot \bigl[1 + \log(B_T \cdot T/H)\bigr] \\ &\le 2C_N \cdot \gamma^2 + C' \cdot \gamma \cdot \log(\gamma T) + C_N \cdot \gamma^2 \cdot \log(B_T \cdot T/H), \end{aligned} \tag{B.1}$$

where $C_N > 0$ is the absolute constant given in Lemma D.1 and C' is an absolute constant that depends on C_N and C_K. Thus, setting $B_T = C_b \cdot \gamma H \cdot \sqrt{\log(dTH)}$ in (B.1), the left-hand side (LHS) of (4.5) is bounded by

$$\begin{aligned}\text{LHS of (4.5)} &\le 8C_K \cdot \gamma \log T + 16C_N \cdot \gamma^2 + 8C' \cdot \gamma \cdot \log(\gamma T) + \\ &\qquad 8C_N \cdot \gamma^2 \cdot \log(C_b \cdot \gamma T \cdot \sqrt{\log(dTH)}) + 16 \cdot \log(TH) + 22 + 2R_Q^2 \\ &\le \gamma^2 \cdot \bigl[\overline{C}_1 \cdot \log(\gamma TH) + 8C_N \cdot \log(C_b)\bigr], \end{aligned} \tag{B.2}$$

where $\overline{C}_1$ is an absolute constant that depends on C', C_N, C_K, and R_Q. Thus, setting C_b as a sufficiently large constant, by (B.2), we have

$$\text{LHS of (4.5)} \le C_b^2 \cdot \gamma^2 \cdot \log(dTH) = (B_T/H)^2,$$

which establishes (4.5) for the first case. Thus, applying Theorem 4.2 we obtain that

$$\text{Regret}(T) \le 8B_T \cdot H \cdot \sqrt{T \cdot \Gamma_K(T,\lambda)} \le C_{r,1} \cdot H^2 \cdot \sqrt{\gamma^3 T} \cdot \log(\gamma TH) = \widetilde{\mathcal{O}}(H^2\sqrt{\gamma^3 T})$$

holds with probability at least $1 - (T^2H^2)^{-1}$, where $C_{r,1}$ is an absolute constant and $\widetilde{\mathcal{O}}(\cdot)$ omits the logarithmic factor. Therefore, we conclude the first case.

Case (ii): γ-Exponential Decay. For the second case, by Lemma D.5 we have

$$\Gamma_K(T,\lambda) \le C_K \cdot (\log T)^{1+1/\gamma}, \tag{B.3}$$

where C_K is an absolute constant. Thus, by the choice of B_T in (4.8), when C_b is sufficiently large, it holds that $R_T = 2H\sqrt{\Gamma_K(T,\lambda)} \le B_T$. Then by Lemma D.1 we have

$$\begin{aligned}\log N_\infty(h, \epsilon^*, B_T) &\le C_N \cdot \bigl[1 + \log(R_T/\epsilon^*)\bigr]^{1+1/\gamma} + C_N \cdot \bigl[1 + \log(B_T/\epsilon^*)\bigr]^{1+2/\gamma} \\ &\le 2C_N \cdot \bigl[1 + \log(B_T/\epsilon^*)\bigr]^{1+2/\gamma} = 2C_N \cdot \bigl\{1 + \log\bigl[C_b T \cdot \sqrt{\log(TH)} \cdot (\log T)^{1/\gamma}\bigr]\bigr\}^{1+2/\gamma}, \end{aligned}$$

where the absolute constant C_N is given by Lemma D.1. By direct computation, there exists an absolute constant $\overline{C}_2$ such that

$$\log N_\infty(h, \epsilon^*, B_T) \le 2C_N \cdot \bigl[1 + \log(C_b) + \overline{C}_2 \cdot \log T + 1/2 \cdot \log\log H\bigr]^{1+2/\gamma}. \tag{B.4}$$

Thus, combining (B.3) and (B.4), the left-hand side of (4.5) is bounded by

$$\begin{aligned}\text{LHS of (4.5)} &\le 8C_K \cdot (\log T)^{1+1/\gamma} + 16C \cdot \bigl[1 + \log(C_b) + \overline{C}_2 \cdot \log T + 1/2 \cdot \log\log H\bigr]^{1+2/\gamma} \\ &\qquad + 16 \cdot \log(TH) + 22 + 2R_Q^2 \\ &\le \overline{C}_3 \cdot \bigl[(\log T)^{1+2/\gamma} + (\log\log H)^{1+2/\gamma} + \log(C_b)\bigr], \end{aligned} \tag{B.5}$$

where $\overline{C}_3$ is an absolute constant that does not depend on C_b. Thus, when C_b is sufficiently large, (B.5) implies that

$$\text{LHS of (4.5)} \le \overline{C}_3 \cdot \big[(\log T)^{1+2/\gamma} + (\log\log H)^{1+2/\gamma} + \log(C_b)\big] \le C_b^2 \cdot (\log T)^{2/\gamma} \cdot \log(TH) = (B_T/H)^2.$$

Thus, for the case of γ-exponential eigenvalue decay, (4.5) holds true for B_T defined in (4.8).

Finally, applying Theorem 4.2 and combining (4.8) and (B.3), we obtain that

$$\mathrm{Regret}(T) \le C_{r,2} \cdot H^2 \cdot \log(TH) \cdot \sqrt{(\log T)^{3/\gamma} \cdot T},$$

where $C_{r,2}$ is an absolute constant. Now we conclude the second case.

Case (iii): γ-Polynomial Decay. Finally, it remains to consider the last case where the eigenvalues satisfy the γ-polynomial decay condition. By Lemma D.5, we have

$$\Gamma_K(T, \lambda) \le C_K \cdot T^{(d+1)/(\gamma+d)} \cdot \log T, \tag{B.6}$$

where C_K is an absolute constant. By direct computation, it holds that

$$R_T/\epsilon^* = 2T\sqrt{\Gamma(T,\lambda)} \le 2\sqrt{C_K} \cdot T^{(2d+\gamma+1)/(\gamma+d)} \cdot \log T. \tag{B.7}$$

To simplify the notation, in the sequel, we let $\widetilde{\gamma}$ denote $\gamma(1-2\tau)$. Besides, we write κ^* defined in (4.9) equivalently as $\kappa^* = \max\{\kappa_1, \kappa_2, \kappa_3\}$ for notational simplicity, where

$$\kappa_1 = \frac{d+1}{2(\gamma+d)} = \xi^*, \qquad \kappa_2 = \frac{2d+\gamma+1}{(d+\gamma)\cdot(\widetilde{\gamma}-1)}, \qquad \kappa_3 = \frac{2}{\widetilde{\gamma}-3}. \tag{B.8}$$

With $B_T = C_b \cdot H \log(TH) \cdot T^{\kappa^*}$, it holds that

$$B_T/\epsilon^* = C_b \cdot \log(TH) \cdot T^{1+\kappa^*}. \tag{B.9}$$

Meanwhile, Lemma D.1 implies that

$$\begin{aligned}
&\log N_\infty(h, \epsilon^*, B_T) \qquad\qquad (\text{B.10})\\
&\quad \le C_N \cdot (2\sqrt{C_K})^{2/(\widetilde{\gamma}-1)} \cdot \Big(T^{(2d+\gamma+1)/(\gamma+d)} \cdot \log T\Big)^{2/(\widetilde{\gamma}-1)} \cdot \big[1 + \log(R_T/\epsilon^*)\big] \\
&\qquad + C_N \cdot C_b^{4/(\widetilde{\gamma}-1)} \cdot T^{4(1+\kappa^*)/(\widetilde{\gamma}-1)} \cdot \big[\log(TH)\big]^{4/(\widetilde{\gamma}-1)} \cdot \big[1 + \log(B_T/\epsilon^*)\big] \\
&\quad \le \overline{C}_4 \cdot T^{2\kappa_2} \cdot (\log T)^{1+2/(\widetilde{\gamma}-1)} + \overline{C}_4 \cdot C_b^{4/(\widetilde{\gamma}-1)} \cdot T^{4(1+\kappa^*)/(\widetilde{\gamma}-1)} \cdot \log(C_b \cdot TH) \cdot \big[\log(TH)\big]^{4/(\widetilde{\gamma}-1)},
\end{aligned}$$

where $\overline{C}_4$ is an absolute constant that only depends on C_N and C_K. Here in the first inequality of (B.10) we plug in (B.7) and (B.9) while in the second inequality we utilize the definition of κ_2 in (B.8) and

$$\log(R_T/\epsilon^*) \asymp \log T, \qquad \log(B_T/\epsilon^*) \asymp \log(C_b) + \log T + \log\log H \le \log(C_b \cdot TH).$$

Moreover, since $\kappa < 1/2$, by the last equality in (B.8), it holds that $2/(\widetilde{\gamma}-1) < 1/3$. Since $\kappa^* \geq \kappa_3$, we have $4(1+\kappa^*)/(\widetilde{\gamma}-1) \leq 2\kappa^*$. Besides, when C_b is sufficiently large, it holds that $\log(C_b) < C_b^{1/3}$, which implies that $C_b^{4/(\widetilde{\gamma}-1)} \cdot \log(C_b) \leq C_b$. Thus, (B.10) can be further simplified into

$$\log N_\infty(h, \epsilon^*, B_T) \leq \overline{C}_4 \cdot T^{2\kappa_2} \cdot (\log T)^2 + \overline{C}_4 \cdot C_b \cdot T^{2\kappa^*} \cdot \big[\log(TH)\big]^2, \tag{B.11}$$

Finally, combining (B.6) and (B.11), the left-hand side of (4.5) is bounded by

$$\begin{aligned}\text{LHS of (4.5)} &\leq \overline{C}_5 \cdot T^{2\kappa_1} \cdot \log T + \overline{C}_5 \cdot T^{2\kappa_2} \cdot (\log T)^2 + \overline{C}_5 \cdot C_b \cdot T^{2\kappa^*} \cdot \big[\log(TH)\big]^2 \\ &\leq C_b^2 \cdot \big[\log(TH)\big]^2 \cdot T^{2\kappa^*} = (B_T/H)^2,\end{aligned}$$

where $\overline{C}_5$ is an absolute constant and the last inequality holds when C_b is sufficiently large. Thus, we establish (4.5). Combining (4.16) in Theorem 4.2, (4.8), and (B.6), the regret of KOVI under this case is bounded by

$$\text{Regret}(T) \leq C_{r,3} \cdot H^2 \cdot T^{\kappa^*+\xi^*+1/2} \cdot \big[\log(TH)\big]^{3/2} = \widetilde{\mathcal{O}}\big(H^2 \cdot T^{\kappa^*+\xi^*+1/2}\big),$$

where $C_{r,3}$ is an absolute constant and $\widetilde{\mathcal{O}}(\cdot)$ omits the logarithmic factor. Thus, we establish the last inequality in (4.16). Therefore, we conclude the proof of Corollary 4.4. □

B.2 Proof of Corollary 4.8

Proof. By Theorem 4.6, we have

$$\text{Regret}(T) = 5\beta H \cdot \sqrt{T \cdot \Gamma_{K_m}(T, \lambda)} + 10\beta T H \cdot \iota \tag{B.12}$$

where $\beta = B_T$ satisfies (4.13) and $\iota = T^{7/12} \cdot H^{1/6} \cdot m^{-1/12} \cdot (\log m)^{1/4}$. When Assumption 4.7 holds, thanks to the similarity between (4.5) and (4.13), it can be similarly shown that B_T defined in (4.8) satisfies the inequality in (4.13) when C_b is sufficiently large. Moreover, Lemma D.5 provides upper bounds on $\Gamma_{K_m}(T, \lambda)$ for all the three eigenvalue decay conditions. Finally, combining (4.8), (B.12), and Lemma D.5, we conclude the proof of Corollary 4.8. □

B.3 Examples of Kernels Satisfying Assumption 4.3

In the following, we introduce concrete kernels and neural tangent kernels that satisfy Assumption 4.3. We consider each eigenvalue decay condition separately.

Case (i): γ-Finite Spectrum. Consider the polynomial kernel $K(z, z') = (1 + \langle z, z' \rangle)^n$ defined on the unit ball $\{z \in \mathbb{R}^d : \|z\|_2 \leq 1\}$, where n is a fixed number. By direct computation, the kernel function can be written as

$$K(z, z') = \sum_{\alpha:\, \|\alpha\|_1 \leq n} z^\alpha \cdot z'^\alpha,$$

where we $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$ is a multi-index and z^α is a monomial with degree α. It can be shown that all monomials in $\mathbb{R}^d$ with degree no more than n are linearly independent. Thus, the dimension of such an RKHS is $\binom{n+d}{d}$, i.e., it satisfy the γ-finite spectrum condition with $\gamma = \binom{n+d}{d}$.

Furthermore, for a finite-dimensional NTK, we consider the quadratic activation function $\mathrm{act}(u) = u^2$. Note that we assume $\mathcal{Z} = \mathbb{S}^{d-1}$ for the neural setting. Moreover, in (2.5), instead of sampling $W_j \sim N(0, I_d/d)$ for all $j \in [d]$, we draw W_j uniformly over the unit sphere $\mathbb{S}^{d-1}$. Then it holds that $|W_j^\top z| \leq 1$ for all $j \in [2m]$ and $z \in \mathbb{S}^{d-1}$. Here we let the distribution be $\mathrm{Unif}(\mathbb{S}^{d-1})$ in order to ensure that the act' is Lipschitz continuous on $\{W_j^\top z : z \in \mathbb{S}^{d-1}\} \subseteq [-1, 1]$ for any W_j sampled from the initial distribution, which is required when utilizing Proposition C.1 in Gao et al. (2019) in the proof of Lemma 5.4. Note that the covariance of W_j is still I_d/d. Then by (2.9), the NTK is given by

$$K_{\mathrm{ntk}}(z, z') = \mathbb{E}_{w \sim \mathrm{Unif}(\mathbb{S}^{d-1})}[2(w^\top z) \cdot 2(w^\top z') \cdot (z^\top z')] = 4/d \cdot (z^\top z')^2, \qquad \forall z, z' \in \mathbb{S}^{d-1}. \tag{B.13}$$

Thus, $K_{\mathrm{ntk}}(z, z')$ can be written as a univariate function of the inner product $\langle z, z' \rangle$. To characterize the spectral property K_{ntk}, we first introduce some background on spherical harmonic functions on $\mathbb{S}^{d-1}$, which are closely related to inner product kernels on $\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$.

Let μ be the uniform measure on $\mathbb{S}^{d-1}$. For any $j \geq 0$, let $\mathcal{Y}_j(d)$ be the set of all homogeneous harmonics of degree j on $\mathbb{S}^{d-1}$, which is a finite-dimensional subspace of $\mathcal{L}^2_\mu(\mathbb{S}^{d-1})$, the space of square-integrable functions on $\mathbb{S}^{d-1}$ with respect to μ. It can be shown that the dimensionality of $\mathcal{Y}_j(d)$ is given by $N(d, j)$, which is defined as

$$N(d, j) = \frac{(2j + d - 2)(d + j - 3)!}{j!(d-2)!}. \tag{B.14}$$

In addition, let $\{Y_{j,\ell}\}_{\ell \in [N(d,j)]}$ be an orthonormal basis of $\mathcal{Y}_j(d)$, then $\{Y_{j,\ell}\}_{\ell \in [N(d,j)], j \in \mathbb{N}}$ form an orthonormal basis of $\mathcal{L}^2_\mu(\mathbb{S}^{d-1})$. In the next lemma, we present the Funk-Hecke formula (Müller, 2012, page 30), which relates spherical harmonics to inner product kernels.

Lemma B.1 (Funk-Hecke formula). Let $k : [-1, 1] \to \mathbb{R}$ be a continuous function, which give rise to an inner product kernel $K(z, z') = k(\langle z, z' \rangle)$ on $\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$. For any $\ell \geq 2$, let $|\mathbb{S}^{\ell-1}|$ be the Lebesgue measure of $\mathbb{S}^{\ell-1}$, which is given by $|\mathbb{S}^{\ell-1}| = 2\pi^{\ell/2}/\Gamma(\ell/2)$, where $\Gamma(\cdot)$ is the Gamma function. Besides, for any $j \geq 0$, let $Y_j : \mathbb{S}^{d-1} \to \mathbb{R}$ be any function in $\mathcal{Y}_j(d)$. Then for any $z \in \mathbb{S}^{d-1}$, we have

$$\int_{\mathbb{S}^{d-1}} K(z, z') Y_j(z') \,\mathrm{d}\mu(z') = \left[\frac{|\mathbb{S}^{d-2}|}{|\mathbb{S}^{d-1}|} \cdot \int_{-1}^{1} k(u) \cdot P_j(u; d) \cdot (1 - u^2)^{(d-3)/2} \,\mathrm{d}u\right] \cdot Y_j(z), \tag{B.15}$$

where $P_j(\cdot; d)$ is the j-th Legendre polynomial in dimension d, which is given by

$$P_j(u; d) = \frac{(-1/2)^j \cdot \Gamma(\frac{d-1}{2})}{\Gamma(\frac{2j+d-1}{2})} \cdot (1 - u^2)^{(3-d)/2} \cdot \left(\frac{\mathrm{d}}{\mathrm{d}u}\right)^j \left[(1 - u^2)^{j + (d-3)/2}\right].$$

Thus, by the Funk-Hecke formula, for any inner product kernel K, its integral operator $T_K : \mathcal{L}^2_\mu(\mathbb{S}^{d-1}) \to \mathcal{L}^2_\mu(\mathbb{S}^{d-1})$ has eigenvalues

$$\varrho_j = \frac{|\mathbb{S}^{d-2}|}{|\mathbb{S}^{d-1}|} \cdot \int_{-1}^{1} k(u) \cdot P_j(u; d) \cdot (1 - u^2)^{(d-3)/2} \,\mathrm{d}u, \qquad \forall j \geq 1, \tag{B.16}$$

each with multiplicity $N(d, j)$. Besides, for each eigenvalue ϱ_j, the corresponding eigenfunctions are spherical harmonics $\{Y_{j,\ell}\}_{\ell \in [N(d,j)]}$. Furthermore, to compute the eigenvalues in (B.16), we have the Rodrigues' rule (Müller, 2012, page 23) as follows.

Lemma B.2 (Rodrigues' Rule)**.** For any $j \geq 0$, let $f\colon [-1,1] \to \mathbb{R}$ be any j-th continuously differentiable function. Then we have

$$\int_{-1}^{1} f(t) \cdot P_j(u;d) \cdot (1-u^2)^{(d-3)/2} \, \mathrm{d}u = R_j(d) \cdot \int_{-1}^{1} f^{(j)}(u) \cdot (1-u^2)^{(2j+d-3)/2} \, \mathrm{d}t,$$

where $f^{(j)}$ is the j-th order derivative of f and $R_j(d) = 2^{-j} \cdot \Gamma((d-1)/2) \cdot [\Gamma((2j+d-1)/2)]^{-1}$ is the j-th Rodrigues constant.

Now we consider the NTK given in (B.13), which is the inner product kernel induced by univariate function $k_1(u) = 4/d \cdot u^2$. Note that $k_1^{(3)}$ is a zero function. Combing Lemma B.2 and (B.16), we observe that $\varrho_j = 0$ for all $j \geq 3$. In addition, by direct computation, it holds that

$$\varrho_1 = R_1(d) \cdot (8/d) \cdot \int_{-1}^{1} u \cdot (1-u^2)^{(d-1)/2} \, \mathrm{d}u = 0$$

and that $\varrho_0, \varrho_2 > 0$. Thus, K_{ntk} given in (B.13) has $N(d,0) + N(d,2) = d(d+1)/2$ nonzero eigenvalues, each with value ϱ_2. This implies that the NTK induced by neural networks with quadratic activation satisfies γ-finite spectrum condition with $\gamma = d(d+1)/2$. For such a class neural networks, Corollary 4.8 asserts that the regret of NOVI is $\widetilde{\mathcal{O}}(H^2 d^3 \cdot \sqrt{T} + \beta T H \cdot \iota)$.

Case (ii): γ-exponential Decay. Now we consider the squared exponential kernel

$$K(z,z') = \exp(-\|z-z'\|_2^2 \cdot \sigma^{-2}) = k_2(\langle z, z'\rangle), \qquad \forall z, z' \in \mathbb{S}^{d-1}, \tag{B.17}$$

where $\sigma > 0$ is an absolute constant and we define $k_2(u) = \exp[-2\sigma^{-2} \cdot (1-u)]$. Note that d is regarded as a fixed number. Applying Lemmas B.1 and B.2, we obtain the following lemma that bounds the eigenvalues of T_K.

Lemma B.3 (Theorem 2 in (Minh et al., 2006))**.** For the squared quadratic kernel in (B.17), the corresponding integral operator has eigenvalues $\{\rho_j\}_{j\geq 0}$, where each ρ_j is defined in (B.16) with k replaced by k_2. Besides, each ϱ_j has multiplicity $N(d,j)$ and the corresponding eigenfunctions are $\{Y_{j,\ell}\}_{\ell\in[N(d,j)]}$. Moreover, when σ in (B.17) satisfy $\sigma^2 \geq 2/d$, $\{\varrho_j\}_{j\geq 0}$ form a decreasing sequence that satisfy

$$A_1 \cdot (2e/\sigma^2)^j \cdot (2j+d-2)^{-(2j+d-1)/2} < \varrho_j < A_2 \cdot (2e/\sigma^2)^j \cdot (2j+d-2)^{-(2j+d-1)/2} \tag{B.18}$$

for all $j \geq 0$, where A_1, A_2 are absolute constants that only depend on d and σ.

Furthermore, the ℓ_∞-norm of each eigenfunction $Y_{j,\ell}$ is given by the following lemma.

Lemma B.4 (Lemma 3 in Minh et al. (2006))**.** For any $d \geq 2$, $j \geq 0$, and any $\ell \in [N(d,j)]$, we have

$$\|Y_{j,\ell}\|_\infty = \sup_{z\in\mathbb{S}^{d-1}} |Y_{j,\ell}(z)| \leq \sqrt{N(d,j)/|\mathbb{S}^{d-1}|}.$$

Now, let $\tau > 0$ be a sufficiently small constant. Combining Lemmas B.3 and B.4, we have

$$\varrho_j^\tau \cdot \|Y_{j,\ell}\|_\infty \leq C \cdot \Big(\frac{2e}{\sigma^2 \cdot (2j+d-2)}\Big)^{-j\cdot\tau} \cdot \sqrt{N(d,j) \cdot (2j+d-2)^{-(d-1)\cdot\tau}}, \tag{B.19}$$

where C is a constant depending on d and σ. By the definition of $N(d,j)$ in (B.14), when j is sufficiently large, it holds that

$$N(d,j) \asymp \frac{(2j+d-2) \cdot \sqrt{d+j-3} \cdot [(d+j-3)/e]^{d+j-3}}{\sqrt{j} \cdot (j/e)^j} \asymp j^{d-2}, \tag{B.20}$$

where we utilize the Stirling's formula and neglect constants involving d. Then combining (B.19) and (B.20) we have

$$\sup_{j\geq 0} \sup_{\ell \in [N(d,j)]} \varrho_j^\tau \cdot \|Y_{j,\ell}\|_\infty \leq C_\varrho \tag{B.21}$$

for some absolute constant $C_\varrho > 0$. Renaming the eigenvalues and eigenvectors as $\{\sigma_j, \psi_j\}_{j\geq 1}$ in the descending order of the eigenvalues, (B.21) equivalently states that $\sup_{j\geq 1} \sigma_j^\tau \cdot \|\psi_j\|_\infty \leq C_\varrho$.

Furthermore, to show that the squared exponential kernel satisfy the γ-exponential decay condition, we notice that

$$\sigma_j = \varrho_t \qquad \text{for} \quad \sum_{i=1}^{t-1} N(d,i) \leq j < \sum_{i=1}^{t} N(d,i). \tag{B.22}$$

Then by (B.20), this implies that $\sigma_j \asymp \rho_t$ for $(t-1)^{d-1} \leq j \leq t^{d-1}$ when j is sufficiently large. Thus, by Lemma B.3 we further obtain that

$$\sigma_j \asymp (2e/\sigma^2)^{j^{\frac{1}{d-1}}} \cdot (2j^{\frac{1}{d-1}} + d - 2)^{-j^{\frac{1}{d-1}} - (d-1)/2} \asymp \exp\big(c_1 \cdot j^{\frac{1}{d-1}}\big) \cdot \exp\big(c_2 - j^{\frac{1}{d-1}} \cdot \log j\big) \leq \exp(-c \cdot j^{1/d}),$$

where c, c_1, and c_2 are constants depending on d. Therefore, we have shown that the squared exponential kernel satisfy the γ-exponential decay condition with $\gamma = 1/d$. Combining this with (B.21), we conclude that it satisfies Assumption 4.3.

Furthermore, in the sequel, we construct an NTK that satisfy Assumption 4.3. Specifically, we adopt the sine activation function and slightly modify the neural network in (2.5) by adding an intercept in each neuron. that is,

$$f(z; b, W, \theta) = \frac{1}{\sqrt{m}} \sum_{j=1}^m b_j \cdot \sin(W_j^\top z + \theta_j).$$

To initialize the network weights (b, W, θ), we set $b_j = -b_{j-m}$, $W_j = W_{j-m}$, and $\theta_j = \theta_{j-m}$ for any $j \in \{m+1, \ldots, 2m\}$. For any $j \in [m]$, we independently sample $b_j \sim \mathrm{Unif}(\{-1,1\})$, $W_j \sim N(0, I_d)$, and $\theta_j \sim \mathrm{Unif}([0, 2\pi])$. Only W is updated during training.

For such a neural network, the corresponding NTK is given by

$$\begin{aligned} K_{\mathrm{ntk}}(z,z') &= 2\mathbb{E}\big[(z^\top z') \cdot \cos(w^\top z + \theta) \cdot \cos(w^\top z' + \theta)\big] \\ &= (z^\top z') \cdot \exp(-\|z - z'\|_2^2/2) = (z^\top z') \cdot \exp[(z^\top z') - 1] = k_3(\langle z, z'\rangle), \end{aligned} \tag{B.23}$$

where we define $k_3(u) = u \cdot \exp(u-1)$. Here the second equality follows from Rahimi and Recht (2008). By construction, such an NTK is closely related to the squared quadratic kernel in (B.17). To see that it satisfy the γ-exponential decay condition, let $\{\varrho_j\}_{j\geq 0}$ and $\{\widetilde{\varrho}_j\}_{j\geq 0}$ denote the eigenvalues of the NTK in (B.23) and the inner product kernel induced by $k_2(u) = \exp(u-1)$, respectively. By Lemma B.1, we have

$$
\begin{aligned}
\rho_j &= C_1 \cdot \int_{-1}^{1} k_3(u) \cdot P_j(u;d) \cdot (1-u^2)^{(d-3)/2} \,\mathrm{d}u = C_1 \cdot \int_{-1}^{1} k_2(u) \cdot u \cdot P_j(u;d) \cdot (1-u^2)^{(d-3)/2} \,\mathrm{d}u \\
&= C_2 \cdot j/(2j+d-2) \cdot \widetilde{\varrho}_{j-1} + C_2 \cdot (j+d-2)/(2j+d-2) \cdot \widetilde{\varrho}_{j+1} \leq C_2(\widetilde{\rho}_{j-1} + \widetilde{\rho}_{j+1}), \qquad \text{(B.24)}
\end{aligned}
$$

where C_1 and C_2 are constants and in the second equality, we utilize the following recurrence relation of Legendre polynomials:

$$
u \cdot P_j(u;d) = j/(2j+d-2) \cdot P_{j-1}(u;d) + (j+d-2)/(2j+d-2) \cdot P_{j+1}(u;d).
$$

Notice that $\{\widetilde{\varrho}_j\}_{j\geq 0}$ satisfy (B.18). Thus, combining (B.18) and (B.24), we obtain (B.21). Moreover, when ordering all the eigenvalues of K_{ntk} in the descending order and renaming them as $\{\sigma_j\}_{j\geq 1}$, similar to (B.22), we have

$$
\sigma_j \leq C_2 \cdot (\widetilde{\rho}_{t-1} + \widetilde{\rho}_{t+1}) \qquad \text{for} \quad \sum_{i=1}^{t-1} N(d,i) \leq j < \sum_{i=1}^{t} N(d,i). \qquad \text{(B.25)}
$$

Using similar analysis, we can show that $\{\sigma_j\}_{j\geq 1}$ satisfy the γ-exponential eigenvalue decay condition with $\gamma = 1/d$. Therefore, we have shown that the NTK given in (B.23) satisfy Assumption 4.3.

Case (iii): γ-Polynomial Decay. Finally, for the last case, it is stated in Srinivas et al. (2009) that the Matérn kernel on $[0,1]^d$ with parameter $\nu > 0$ satisfies the γ-polynomial decay condition with $\gamma = 1 + 2\nu/d$. Moreover, the eigenfunctions are Sinusoidal functions and thus are bounded. Hence, Matérn kernel satisfy the last case of Assumption 4.3 with $\gamma = 1 + 2\nu/d$ and $\tau = 0$. As a result, in this case, (B.8) reduces to

$$
\kappa_1 = \xi^* = \frac{d(d+1)}{2[2\nu + d(d+1)]}, \qquad \kappa_2 = \frac{d(d+1)+\nu}{(d+1+2\nu/d)\nu}, \qquad \kappa_3 = \frac{d}{\nu - d}.
$$

We further assume ν is sufficiently large such that $\Gamma_K(T,\lambda) \cdot \sqrt{T} = o(T)$, which implies that $\kappa_1 < 1/4$ and that $2\nu > d(d+1)$. In this case, we have

$$
\kappa_2 = \frac{1 + d(d+1)/\nu}{d+1+2\nu/d} < \frac{3}{d+1}, \qquad \kappa_2 < \frac{2}{d-1}, \qquad \kappa^* = \max\Bigl\{\xi^*, \frac{3}{d+1}, \frac{2}{d-1}\Bigr\}.
$$

Thus, (4.16) reduces to an $\widetilde{\mathcal{O}}(H^2 \cdot T^{\kappa^* + \xi^* + 1/2})$ regret.

Finally, we construct an NTK that satisfy Assumption 4.3. Similar to the NTK given in (B.13), with $\mathcal{Z} = \mathbb{S}^{d-1}$, for the neural network in (2.5), we let the activation function be $\mathrm{act}(u) = (u)_+^{s+1}$ and set $W_j \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(\mathbb{S}^{d-1})$. Here s is a positive integer and $(u)_+ = u \cdot \mathbb{1}\{u \geq 0\}$ is the ReLU activation function. By direct computation, the induced NTK is

$$
K_{\mathrm{ntk}}(z,z') = K_s(z,z') \cdot (z^\top z'), \qquad K_s(z,z') = (s+1)^2 \cdot \mathbb{E}_{w\sim \mathrm{Unif}(\mathbb{S}^{d-1})}\bigl[(w^\top z)_+^s (w^\top z')_+^s\bigr]. \qquad \text{(B.26)}
$$

Utilizing the rotational invariance of Unif($\mathbb{S}^{d-1}$), it can be shown that K_{ntk} and K_s in (B.26) are both inner-product kernels, i.e., there exist univariate functions $k_4, k_5 \colon [-1,1] \to \mathbb{R}$ such that

$$K_{\text{ntk}}(z,z') = k_4(\langle z, z'\rangle), \qquad K_s(z,z') = k_5(\langle z, z'\rangle), \qquad k_4(u) = u \cdot k_5(u).$$

By Lemma B.1, for any $j \geq 0$. K_{ntk} and K_s both have spherical harmonics $\mathcal{Y}_j(d)$ as their eigenvectors. Let the corresponding eigenvalues be ϱ_j and $\widetilde{\varrho}_j$, respectively. Similar to (B.24), we have $\varrho_j \leq C_2(\widetilde{\varrho}_{j-1} + \widetilde{\rho}_{j+1})$, where C_2 is an absolute constant depending on d. Furthermore, as shown in §D in Bach (2017), when j is sufficiently large, it holds that $\widetilde{\varrho}_j \asymp j^{-(d+2s)}$. This further implies that $\varrho_j = \mathcal{O}(j^{-(d+2s)})$ where $\mathcal{O}(\cdot)$ hides constants depending on d. Thus, for $\tau = (d-2)/(2d+4s)$, combining Lemma B.4 and (B.20), we have

$$\varrho_j^{\tau} \cdot \|Y_{j,\ell}\|_{\infty} \leq C \cdot j^{-(d+2s)\cdot\tau} \cdot j^{(d-2)/2} = C_{\varrho} \cdot j^{-[(2d+4s)\cdot\tau - (d-2)]/2} \leq C_{\varrho} \tag{B.27}$$

when j is sufficiently large, where C_{ϱ} is an absolute constant depending on d. Moreover, we have s $\tau \in [0, 1/2)$ and that τ can be made sufficiently small by increasing s.

Finally, let $\{\sigma_j\}_{j\geq 1}$ be all the eigenvalues of K_{ntk} in descending order. By (B.25) and (B.20), when j is sufficiently large, we have

$$\sigma_j = \mathcal{O}\big((j^{\frac{1}{d-1}})^{-(d+2s)}\big) = \mathcal{O}\big(j^{-(1+\frac{1+2s}{d-1})}\big),$$

where $\mathcal{O}(\cdot)$ hides constants depending on d. Therefore, K_{ntk} in (B.26) satisfy the γ-polynomial decay condition with $\gamma = 1 + (1+2s)/(d-1)$. Combining this with (B.27), we conclude that K_{ntk} satisfy the last case of Assumption 4.3.

C Proofs of the Supporting Lemmas

C.1 Proof of Lemma 5.1

Proof. For ease of presentation, before presenting the proof, we first define two operators $\mathbb{J}_h^{\star}$ and $\mathbb{J}_{t,h}$ respectively by letting

$$(\mathbb{J}_h^{\star} f)(x) = \langle f(x,\cdot), \pi_h^{\star}(\cdot\,|\,x)\rangle, \quad (\mathbb{J}_{t,h} f)(x) = \langle f(x,\cdot), \pi_h^{t}(\cdot\,|\,x)\rangle \tag{C.1}$$

for any $(t,h) \in [T]\times[H]$ and any function $f : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$. Moreover, for any $(t,h) \in [T]\times[H]$ and any state $x \in \mathcal{S}$, we define

$$\xi_h^t(x) = (\mathbb{J}_h Q_h^t)(x) - (\mathbb{J}_{t,h} Q_h^t)(x) = \langle Q_h^t(x,\cdot), \pi_h^{\star}(\cdot\,|\,x) - \pi_h^t(\cdot\,|\,x)\rangle. \tag{C.2}$$

After introducing these notations, to prove (5.4), we decompose the instantaneous regret at the t-th episode into the following two terms,

$$V_1^{\star}(x_1^t) - V_1^{\pi^t}(x_1^t) = \underbrace{V_1^{\star}(x_1^t) - V_1^t(x_1^t)}_{\text{(i)}} + \underbrace{V_1^t(x_1^t) - V_1^{\pi^t}(x_1^t)}_{\text{(ii)}}. \tag{C.3}$$

In the sequel, we consider the two terms in (C.3) separately.

Term (i). By the definitions of the value function $V_h^\star$ in (2.2) and the operator $\mathbb{J}_h^\star$ in (C.1), we have $V_h^\star = \mathbb{J}_h^\star Q_h^\star$. Similarly, for all the algorithms, we have $V_h^t(x) = \langle Q_h^t(x,\cdot), \pi_h^t(\cdot\,|\,x)\rangle$ for all $x \in \mathcal{S}$. Thus, by the definition of $\mathbb{J}_{t,h}$ in (C.1), we have $V_h^t = \mathbb{J}_{t,h} Q_h^t$. Thus, using ξ_h^t defined in (C.2), for any $(t,h) \in [T] \times [H]$, we have

$$
\begin{aligned}
V_h^\star - V_h^t &= \mathbb{J}_h^\star Q_h^\star - \mathbb{J}_{t,h} Q_h^t = \big(\mathbb{J}_h^\star Q_h^\star - \mathbb{J}_h^\star Q_h^t\big) + \big(\mathbb{J}_h^\star Q_h^t - \mathbb{J}_{t,h} Q_h^t\big) \\
&= \mathbb{J}_h^\star (Q_h^\star - Q_h^t) + \xi_h^t,
\end{aligned} \tag{C.4}
$$

where the last equality follows from the definition of ξ_h^t in (C.2) and the fact that $\mathbb{J}_h^\star$ is a linear operator. Moreover, by the definition of the temporal-difference error δ_h^t in (5.1) and the Bellman optimality condition, we have

$$
Q_h^\star - Q_h^t = \big(r_h + \mathbb{P}_h V_{h+1}^\star\big) - \big(r_h + \mathbb{P}_h V_{h+1}^t - \delta_h^t\big) = \mathbb{P}_h (V_{h+1}^\star - V_{h+1}^t) + \delta_h^t. \tag{C.5}
$$

Thus, combining (C.4) and (C.5), we obtain that

$$
V_h^\star - V_h^t = \mathbb{J}_h^\star \mathbb{P}_h (V_{h+1}^\star - V_{h+1}^t) + \mathbb{J}_h^\star \delta_h^t + \xi_h^t, \qquad \forall (t,h) \in [T] \times [H]. \tag{C.6}
$$

Equivalently, for all $x \in \mathcal{S}$, and all $(t,h) \in [T] \times [H]$, we have

$$
\begin{aligned}
V_h^\star(x) - V_h^t(x) =& \mathbb{E}_{a \sim \pi_h^\star(\cdot\,|\,x)} \Big\{ \mathbb{E}\big[V_{h+1}^\star(x_{h+1}) - V_{h+1}^t(x_{h+1}) \,\big|\, x_h = x, a_h = a\big] \Big\} \\
&+ \mathbb{E}_{a \sim \pi_h^\star(\cdot\,|\,x)} \big[\delta_h^t(x,a)\big] + \xi_h^t(x).
\end{aligned}
$$

Then, by recursively applying (C.6) for all $h \in [H]$, we have

$$
V_1^\star - V_1^t = \Big(\prod_{h=1}^H \mathbb{J}_h^\star \mathbb{P}_h\Big)(V_{H+1}^\star - V_{H+1}^k) + \sum_{h=1}^H \Big(\prod_{i=1}^{h-1} \mathbb{J}_i^\star \mathbb{P}_i\Big) \mathbb{J}_h^\star \delta_h^t + \sum_{h=1}^H \Big(\prod_{i=1}^{h-1} \mathbb{J}_i^\star \mathbb{P}_i\Big) \xi_h^t. \tag{C.7}
$$

Furthermore, notice that we have $V_{H+1}^\star = V_{H+1}^k = \mathbf{0}$. Thus, (C.7) can be equivalently written as

$$
V_1^\star(x) - V_1^t(x) = \mathbb{E}_{\pi^\star}\bigg[\sum_{h=1}^H \langle Q_h^t(x_h,\cdot), \pi_h^\star(\cdot\,|\,x_h) - \pi_h^t(\cdot\,|\,x_h)\rangle + \delta_h^t(x_h,a_h) \,\bigg|\, x_1 = x\bigg],
$$

where we utilize the definition of ξ_h^t given in (C.2). Thus, we can write Term (i) on the right-hand side of (C.3) as

$$
\begin{aligned}
V_1^\star(x_t^t) - V_1^t(x_t^t) &= \sum_{h=1}^H \mathbb{E}_{\pi^\star}\big[\langle Q_h^t(x_h,\cdot), \pi_h^\star(\cdot\,|\,x_h) - \pi_h^t(\cdot\,|\,x_h)\rangle \,|\, x_1 = x_t^t\big] \\
&\qquad + \sum_{h=1}^H \mathbb{E}_{\pi^\star}[\delta_h^t(x_h,a_h) \,|\, x_1 = x_t^t], \qquad \forall t \in [T].
\end{aligned} \tag{C.8}
$$

Term (ii). It remains to bound the second term on the right-hand side of (C.3). By the definition of the temporal-difference error δ_h^t in (5.1), for any $(t, h) \in [T] \times [H]$, we have

$$\begin{aligned}
\delta_h^t(x_h^t, a_h^t) &= r_h(x_h^t, a_h^t) + (\mathbb{P}_h V_{h+1}^t)(x_h^t, a_h^t) - Q_h^t(x_h^t, a_h^t) \\
&= \big[r_h(x_h^t, a_h^t) + (\mathbb{P}_h V_{h+1}^t)(x_h^t, a_h^t) - Q_h^{\pi^t}(x_h^t, a_h^t)\big] + \big[Q_h^{\pi^t}(x_h^t, a_h^t) - Q_h^t(x_h^t, a_h^t)\big] \\
&= \big(\mathbb{P}_h V_{h+1}^t - \mathbb{P}_h V_{h+1}^{\pi^t}\big)(x_h^t, a_h^t) + (Q_h^{\pi^t} - Q_h^t)(x_h^t, a_h^t),
\end{aligned} \tag{C.9}$$

where the last equality follows from the Bellman equation (2.1). Morerover, recall that we define $\zeta_{t,h}^1$ and $\zeta_{t,h}^2$ in (5.2) and (5.3), respectively. Thus, from (C.9) we obtain that

$$\begin{aligned}
&V_h^t(x_h^t) - V_h^{\pi^t}(x_h^t) \\
&\quad = V_h^t(x_h^t) - V_h^{\pi^t}(x_h^t) + (Q_h^{\pi^t} - Q_h^t)(x_h^t, a_h^t) + \big(\mathbb{P}_h(V_{h+1}^t - V_{h+1}^{\pi^t})\big)(x_h^t, a_h^t) - \delta_h^t(x_h^t, a_h^t), \\
&\quad = \big(V_h^t - V_h^{\pi^t}\big)(x_h^t) - (Q_h^t - Q_h^{\pi^t})(x_h^t, a_h^t) \\
&\qquad\quad + \big(\mathbb{P}_h(V_{h+1}^t - V_{h+1}^{\pi^t})\big)(x_h^t, a_h^t) - (V_{h+1}^t - V_{h+1}^{\pi^t})(x_{h+1}^t) + (V_{h+1}^t - V_{h+1}^{\pi^t})(x_{h+1}^t) - \delta_h^t(x_h^t, a_h^t) \\
&\quad = \big[V_{h+1}^t(x_{h+1}^t) - V_{h+1}^{\pi^t}(x_{h+1}^t)\big] + \zeta_{t,h}^1 + \zeta_{t,h}^2 - \delta_h^t(x_h^t, a_h^t).
\end{aligned} \tag{C.10}$$

Thus, recursively applying (C.10) for all $h \in [H]$, we obtain that

$$\begin{aligned}
V_1^t(x_1^t) - V_1^{\pi^t}(x_1^t) &= V_{H+1}^t(x_{H+1}^k) - V_{H+1}^{\pi^k, k}(x_{H+1}^t) + \sum_{h=1}^H (\zeta_{t,h}^1 + \zeta_{t,h}^2) - \sum_{h=1}^H \delta_h^t(x_h^t, a_h^t) \\
&= \sum_{h=1}^H (\zeta_{t,h}^1 + \zeta_{t,h}^2) - \sum_{h=1}^H \delta_h^t(x_h^t, a_h^t), \qquad \forall t \in [T],
\end{aligned} \tag{C.11}$$

where the last equality follows from the fact that $V_{H+1}^t(x_{H+1}^t) = V_{H+1}^{\pi^t}(x_{H+1}^t) = 0$. Thus, we have simplified Term (ii) defined in (C.3).

Thus, combining (C.3), (C.8), and (C.11), we obtain that

$$\begin{aligned}
\mathrm{Regret}(T) &= \sum_{t=1}^T \big[V_1^\star(x_1^t) - V_1^{\pi^t}(x_1^t)\big] \\
&= \sum_{t=1}^T \sum_{h=1}^H \mathbb{E}_{\pi^\star}[\delta_h^t(x_h, a_h) \,|\, x_1 = x_t^t] + \sum_{t=1}^T \sum_{h=1}^H (\zeta_{t,h}^1 + \zeta_{t,h}^2) - \sum_{t=1}^T \sum_{h=1}^H \delta_h^t(x_h^t, a_h^t) \\
&\quad + \sum_{t=1}^T \sum_{h=1}^H \mathbb{E}_{\pi^\star}\big[\langle Q_h^t(x_h, \cdot), \pi_h^\star(\cdot \,|\, x_h) - \pi_h^t(\cdot \,|\, x_h)\rangle \,\big|\, x_1 = x_t^t\big].
\end{aligned}$$

Therefore, we conclude the proof of this lemma. ∎

C.2 Proof of Lemma 5.2

Proof. For ease of presentation, we utilize the feature representation induced by the kernel K. Let $\phi\colon \mathcal{Z} \to \mathcal{H}$ be the feature mapping such that $K(z, z') = \langle \phi(z), \phi(z') \rangle_{\mathcal{H}}$. For simplicity, we formally view $\phi(z)$ as a vector and write $\langle\phi(z),\phi(z')\rangle_{\mathcal{H}} = \phi(z)^\top\phi(z')$. Then, any function $f\colon \mathcal{Z}\to\mathbb{R}$ in the RKHS satisfies $f(z) = \langle\phi(z), f\rangle_{\mathcal{H}} = f^\top\phi(z)$. Using the feature representation, we can rewrite the kernel ridge regression in (3.4) as

$$\underset{\theta\in\mathcal{H}}{\text{minimize}}\, L(\theta) = \sum_{\tau=1}^{t-1}\big[r_h(x_h^\tau, a_h^\tau) + V_{h+1}^t(x_{h+1}^\tau) - \langle\phi(x_h^\tau, a_h^\tau), \theta\rangle_{\mathcal{H}}\big]^2 + \lambda\cdot\|\theta\|_{\mathcal{H}}^2. \tag{C.12}$$

We define the feature matrix $\Phi_h^t\colon \mathcal{H}\to\mathbb{R}^{t-1}$ and "covariance matrix" $\Lambda_h^t\colon\mathcal{H}\to\mathcal{H}$ respectively as

$$\Phi_h^t = \big[\phi(z_h^1)^\top, \ldots, \phi(z_h^{t-1})^\top\big]^\top, \qquad \Lambda_h^t = \sum_{\tau=1}^{t-1}\phi(z_h^\tau)\phi(z_h^\tau)^\top + \lambda\cdot I_{\mathcal{H}} = \lambda\cdot I_{\mathcal{H}} + (\Phi_h^t)^\top\Phi_h^t, \tag{C.13}$$

where $I_{\mathcal{H}}$ is the identity mapping on $\mathcal{H}$. Thus, the Gram matrix K_h^t in (3.7) is equal to $\Phi_h^t(\Phi_h^t)^\top$. More specifically, here Λ_h^t is a self-adjoint and positive definite operator. For any $f_1, f_2\in\mathcal{H}$, we denote

$$\Lambda_h^t f_1 = \lambda\cdot f_1 + \sum_{\tau=1}^{t-1}\phi(z_h^\tau)\cdot f_1(x_h^\tau)\in\mathcal{H}, \qquad f_1^\top\Lambda_h^t f_2 = \langle f_1, \Lambda_h^t f\rangle_{\mathcal{H}}.$$

It is not hard to see that all the eigenvalues of Λ_h^t are positive and at least λ. Thus, the inverse operator of Λ_h^t, denoted by $(\Lambda_h^t)^{-1}$, is well-defined, which is also a self-adjoint and positive definite operator on $\mathcal{H}$. Similarly, for any $f_1, f_2\in\mathcal{H}$, we let $f_1^\top(\Lambda_h^t)^{-1}f_2$ denote $\langle f_1, (\Lambda_h^t)^{-1}f_2\rangle_{\mathcal{H}}$. The eigenvalues of $(\Lambda_h^t)^{-1}$ are all bounded in interval $[0, 1/\lambda]$.

In addition, using the feature matrix Φ_h^t defined in (C.13) and y_h^t defined in (3.6), we can write (C.12) as

$$\underset{\theta\in\mathcal{H}}{\text{minimize}}\, L(\theta) = \big\|y_h^t - \Phi_h^t\theta\big\|_2^2 + \lambda\cdot\theta^\top\theta,$$

whose solution is given by $\widehat{\theta}_h^t = (\Lambda_h^t)^{-1}(\Phi_h^t)^\top y_h^t$. and $\widehat{Q}_h^t$ in (3.4) satisfies $\widehat{Q}_h^t(z) = \phi(z)^\top\widehat{\theta}_h^t$.

In the sequel, to further simplify the notation, we let Φ denote Φ_h^t when its meaning is clear from the context. Since both $(\Phi\Phi^\top + \lambda\cdot I)$ and $(\Phi^\top\Phi + \lambda\cdot I_{\mathcal{H}})$ are strictly positive definite and

$$(\Phi^\top\Phi + \lambda\cdot I_{\mathcal{H}})\Phi^\top = \Phi^\top(\Phi\Phi^\top + \lambda\cdot I),$$

which implies that

$$(\Lambda_h^t)^{-1}\Phi^\top = (\Phi\Phi^\top + \lambda\cdot I_{\mathcal{H}})^{-1}\Phi^\top = \Phi^\top(\Phi\Phi^\top + \lambda\cdot I)^{-1} = \Phi^\top(K_h^t + \lambda\cdot I)^{-1}. \tag{C.14}$$

Here I is the identity matrix in $\mathbb{R}^{(t-1)\times(t-1)}$. Thus, by (C.14) we have

$$\widehat{\theta}_h^t = (\Lambda_h^t)^{-1}\Phi^\top y_h^t = \Phi^\top(K_h^t + \lambda\cdot I)^{-1}y_h^t = \Phi^\top\alpha_h^t. \tag{C.15}$$

Moreover, k_h^t defined in (3.7) can be written as $k_h^t(z) = \Phi\phi(z)$, which, combined with (C.14), implies

$$\begin{aligned}\phi(z) &= (\Lambda_h^t)^{-1}\Lambda_h^t\phi(z) = (\Lambda_h^t)^{-1}(\Phi^\top\Phi + \lambda\cdot I_{\mathcal{H}})\phi(z)\\ &= (\Lambda_h^t)^{-1}(\Phi^\top\Phi)\phi(z) + \lambda\cdot(\Lambda_h^t)^{-1}\phi(z)\\ &= \Phi^\top(K_h^t + \lambda\cdot I)^{-1}k_h^t(z) + \lambda\cdot(\Lambda_h^t)^{-1}\phi(z).\end{aligned} \tag{C.16}$$

Thus, we can write $\|\phi(z)\|_{\mathcal{H}}^2 = \phi(z)^\top \phi(z)$ as

$$
\begin{aligned}
\|\phi(z)\|_{\mathcal{H}}^2 &= \phi(z)^\top \cdot \big[\Phi^\top (K_h^t + \lambda \cdot I)^{-1} k_h^t(z) + \lambda \cdot (\Lambda_h^t)^{-1}\phi(z)\big] \\
&= k_h^t(z)^\top (K_h^t + \lambda \cdot I)^{-1} k_h^t(z) + \lambda \cdot \phi(z)(\Lambda_h^t)^{-1}\phi(z),
\end{aligned}
$$

which implies that we can equivalently write the bonus b_h^t defined in (3.8) as

$$
b_h^t(x,a) = \big[\phi(x,a)^\top (\Lambda_h^t)^{-1}\phi(x,a)\big]^{1/2} = \|\phi(x,a)\|_{(\Lambda_h^t)^{-1}}. \tag{C.17}
$$

Combining (C.15) and (C.17), we equivalently write Q_h^t in (3.5) as

$$
\begin{aligned}
Q_h^t(x,a) &= \min\big\{\widehat{Q}_h^t(x,a) + \beta \cdot b_h^t(x,a), H-h+1\big\}^+ \\
&= \min\big\{\phi(x,a)^\top \widehat{\theta}_h^t + \beta \cdot \|\phi(x,a)\|_{(\Lambda_h^t)^{-1}}, H-h+1\big\}^+.
\end{aligned} \tag{C.18}
$$

Now we are ready to bound the temporal-difference error ξ_h^t defined in (5.1). Noice that $V_h^t(x) = \max_a Q_h^t(x,a)$ for all $(t,h) \in [T]\times[H]$, we have

$$
\delta_h^t = r_h + \mathbb{P}_h V_{h+1}^t - Q_h^t = \mathbb{T}_h^\star Q_{h+1}^t - Q_h^t,
$$

where $\mathbb{T}_h^\star$ is the Bellman optimality operator. Under the Assumption 4.1, for all $(t,h)\in[T]\times[H]$, since $Q_{h+1}^t \in [0,H]$, we have $\mathbb{T}_h^\star Q_{h+1}^t \in \mathcal{Q}^\star$. Using the feature representation of RKHS, there exists $\overline{\theta}_h^t \in \mathcal{Q}^\star$ such that $(\mathbb{T}_h^\star Q_{h+1}^t)(z) = \phi(z)^\top \overline{\theta}_h^t$ for all $z \in \mathcal{Z}$.

In the sequel, we consider the difference between $\phi(z)^\top\widehat{\theta}_h^t$ and $\phi(z)^\top\overline{\theta}_h^t$. To begin with, using (C.16), we can write $\phi(z)^\top\overline{\theta}_h^t$ as

$$
\phi(z)^\top \overline{\theta}_h^t = k_h^t(z)^\top (K_h^t + \lambda \cdot I)^{-1}\Phi\overline{\theta}_h^t + \lambda \cdot \phi(z)^\top (\Lambda_h^t)^{-1}\overline{\theta}_h^t. \tag{C.19}
$$

Hence, combining (C.15) and (C.19), we have

$$
\phi(z)^\top \widehat{\theta}_h^t - \phi(z)^\top \overline{\theta}_h^t = \underbrace{k_h^t(z)^\top (K_h^t + \lambda \cdot I)^{-1}\big(y_h^t - \Phi\overline{\theta}_h^t\big)}_{\text{(i)}} - \underbrace{\lambda \cdot \phi(z)^\top (\Lambda_h^t)^{-1}\overline{\theta}_h^t}_{\text{(ii)}}. \tag{C.20}
$$

We bound Term (i) and Term (ii) on the right-hand side of (C.20) separately. For Term (ii), by Cauchy-Schwarz inequality, we have

$$
\begin{aligned}
\big|\lambda \cdot \phi(z)^\top (\Lambda_h^t)^{-1}\overline{\theta}_h^t\big| &\le \big\|\lambda \cdot (\Lambda_h^t)^{-1}\phi(x)\big\|_{\mathcal{H}} \cdot \|\overline{\theta}_h^t\|_{\mathcal{H}} \le R_Q H \cdot \big\|\lambda \cdot (\Lambda_h^t)^{-1}\phi(x)\big\|_{\mathcal{H}} \\
&= R_Q H \cdot \sqrt{\lambda \cdot \phi(z)^\top (\Lambda_h^t)^{-1} \cdot \lambda \cdot I_{\mathcal{H}} \cdot (\Lambda_h^t)^{-1}\phi(x)} \\
&\le R_Q H \cdot \sqrt{\lambda \cdot \phi(z)(\Lambda_h^t)^{-1} \cdot \Lambda_h^t \cdot (\Lambda_h^t)^{-1}\phi(z)} = \sqrt{\lambda} R_Q H \cdot b_h^t(z).
\end{aligned} \tag{C.21}
$$

Here the first inequality follows from Cauchy-Schwarz inequality and the second inequality follows from the fact that $\overline{\theta}_h^t \in \mathcal{Q}^\star$, which implies that $\|\overline{\theta}_h^t\|_{\mathcal{H}} \le R_Q H$. Besides, the last inequality follows from the fact that $\Lambda_h^t - \lambda \cdot I_{\mathcal{H}}$ is a self-adjoint and positive-semidefinite operator, which means that $f^\top(\Lambda_h^t - \lambda \cdot I_{\mathcal{H}})f \ge 0$ for all $f \in \mathcal{H}$, and the last equality follows from (C.17).

Furthermore, for Term (i), by the Bellman equation in (2.2) and the definition of y_h^t in (3.6), for any $\tau \in [t-1]$, the τ-th entry of $(y_h^t - \Phi\overline{\theta}_h^t)$ can be written as

$$
\begin{aligned}
[y_h^t]_\tau - [\Phi\overline{\theta}_h^t]_\tau &= r_h(x_h^\tau, a_h^\tau) + V_{h+1}^t(x_{h+1}^\tau) - \phi(x_h^\tau, a_h^\tau)^\top \overline{\theta}_h^t = r_h(x_h^\tau, a_h^\tau) + V_{h+1}^t(x_{h+1}^\tau) - (\mathbb{T}_h^\star Q_{h+1}^t)(x_h^\tau, a_h^\tau) \\
&= V_{h+1}^t(x_{h+1}^\tau) - (\mathbb{P}_h V_{h+1}^t)(x_h^\tau, a_h^\tau). \qquad \text{(C.22)}
\end{aligned}
$$

Thus, combining (C.14), (C.20), and (C.22) we have

$$
\begin{aligned}
&\big|k_h^t(z)^\top (K_h^t + \lambda \cdot I)^{-1} \big(y_h^t - \Phi\overline{\theta}_h^t\big)\big| \\
&\quad = \bigg|\phi(z)^\top (\Lambda_h^t)^{-1} \bigg\{ \sum_{\tau=1}^{t-1} \phi(x_h^\tau, a_h^\tau) \cdot \big[V_{h+1}^k(x_{h+1}^\tau) - (\mathbb{P}_h V_{h+1}^k)(x_h^\tau, a_h^\tau)\big] \bigg\}\bigg| \\
&\quad \le \|\phi(z)\|_{(\Lambda_h^t)^{-1}} \cdot \bigg\| \sum_{\tau=1}^{t-1} \phi(x_h^\tau, a_h^\tau) \cdot \big[V_{h+1}^t(x_{h+1}^\tau) - (\mathbb{P}_h V_{h+1}^t)(x_h^\tau, a_h^\tau)\big] \bigg\|_{(\Lambda_h^t)^{-1}}, \qquad \text{(C.23)}
\end{aligned}
$$

where the last inequality follows from the Cauchy-Schwarz inequality. In the following, we aim to bound (C.23) by the concentration of self-normalized stochastic process in the RKHS. However, here V_{h+1}^t depends on the historical data in the first $(t-1)$ episodes and is thus not independent of $\{(x_h^\tau, a_h^\tau, x_{h+1}^\tau)\}_{\tau \in [t-1]}$. To bypass this challenge, in the sequel, we combine the concentration of self-normalized process and uniform convergence over the function classes that contain each V_{h+1}^t.

Specifically, recall that we define function class $\mathcal{Q}_{\rm ucb}(h, R, B)$ in (4.4) for any $h \in [H]$, and any $R, B > 0$. We define $\mathcal{V}_{\rm ucb}(h, R, B)$ as

$$
\mathcal{V}_{\rm ucb}(h, R, B) = \big\{ V \colon V(\cdot) = \max_{a \in \mathcal{A}} Q(\cdot, a) \ \text{ for some } \ Q \in \mathcal{Q}_{\rm ucb}(h, R, B) \big\}. \qquad \text{(C.24)}
$$

In the following, we find a parameter R_T such that $V_h^t \in \mathcal{V}_{\rm ucb}(h, R_T, B_T)$ holds for all $h \in [H]$ and $t \in [T]$, where B_T is specified in (4.5). Here both R_T and B_T depend on T. By (4.4) and (C.18), it suffices to set R_T as an upper bound of $\|\widehat{\theta}_h^t\|_{\mathcal{H}}$ for all $(t, h) \in [T] \times [H]$. In the following lemma, bound the RKHS norm of each $\widehat{\theta}_h^t$.

Lemma C.1 (Bound on Weights in Algorithm)**.** When $\lambda \ge 1$, for any $(t, h) \in [T] \times [H]$, $\widehat{\theta}_h^t$ defined in (C.15) satisfies

$$
\|\widehat{\theta}_h^t\|_{\mathcal{H}} \le H\sqrt{2/\lambda \cdot \operatorname{logdet}(I + K_h^t/\lambda)} \le 2H\sqrt{\Gamma_K(T, \lambda)},
$$

where K_h^t is defined in (3.7) and $\Gamma_K(T, \lambda)$ is defined in (D.15).

Proof. See §E.1 for a detailed proof. □

By this lemma, in the sequel, we set $R_T = 2H\sqrt{\Gamma_K(T, \lambda)}$. To conclude the proof, we show that the sum of the two terms in (C.20) is bounded by $\beta \cdot \|\phi(z)\|_{(\Lambda_h^t)^{-1}}$, where we set $\beta = B_T$. To this end, for any two value functions $V, V' \colon \mathcal{S} \to \mathbb{R}$, we define their distance as $\operatorname{dist}(V, V') = \sup_{x \in \mathcal{S}} |V(x) - V'(x)|$. For any $\epsilon \in (0, 1/e)$, any $B > 0$, and any $h \in [H]$, we let $N_{\rm dist}(\epsilon; h, B)$ be the ϵ-covering number of $\mathcal{V}_{\rm ucb}(h, R_T, B)$ with respect to distance $\operatorname{dist}(\cdot, \cdot)$. Recall that we define

$N_\infty(\epsilon; h, B)$ as the ϵ-covering number of $\mathcal{Q}_{\rm ucb}(h, R_T, B)$ with respect to the ℓ_∞-norm on $\mathcal{Z}$. Note that for any $Q, Q' \colon \mathcal{Z} \to \mathbb{R}$, we have

$$\sup_{x\in\mathcal{S}}\Big|\max_{a\in\mathcal{A}} Q(x,a) - \max_{a\in\mathcal{A}} Q'(x,a)\Big| \le \sup_{(x,a)\in\mathcal{S}\times\mathcal{A}} |Q(x,a) - Q'(x,a)| = \|Q - Q'\|_\infty.$$

By (C.24) we have $N_{\rm dist}(\epsilon; h, B) \le N_\infty(\epsilon; h, B)$. Then, by applying Lemma E.2 with $\delta = (2T^2H^3)^{-1}$ and taking a union bound over $h \in [H]$, we obtain that

$$\begin{aligned}
&\bigg\|\sum_{\tau=1}^{t-1}\phi(x_h^\tau, a_h^\tau)\cdot\big[V_{h+1}^t(x_{h+1}^\tau) - (\mathbb{P}_h V_{h+1}^t)(x_h^\tau, a_h^\tau)\big]\bigg\|^2_{(\Lambda_h^t)^{-1}} \\
&\quad\le \sup_{V\in\mathcal{V}_{\rm ucb}(h+1,R_T,B_T)}\bigg\|\sum_{\tau=1}^{t-1}\phi(x_h^\tau, a_h^\tau)\cdot\big[V(x_{h+1}^\tau) - (\mathbb{P}_h V)(x_h^\tau, a_h^\tau)\big]\bigg\|^2_{(\Lambda_h^t)^{-1}} \\
&\quad\le 2H^2\cdot\mathrm{logdet}(I + K_h^t/\lambda) + 2H^2 t\cdot(\lambda - 1) + 8t^2\epsilon^2/\lambda \\
&\qquad + 4H^2\cdot\big[\log N_\infty(\epsilon; h+1, B_T) + \log(2T^2H^3)\big],
\end{aligned} \tag{C.25}$$

holds uniformly for all $(t, h) \in [T] \times [H]$ with probability at least $1 - (2T^2H^2)^{-2}$, where we utilize the fact that $V_{h+1}^t \in \mathcal{V}_{\rm ucb}(h+1, R_T, B_T)$. Note that we set $\lambda = 1 + 1/T$. Then, setting ϵ as $\epsilon^* = H/T$, (C.25) is further reduced to

$$\begin{aligned}
&\bigg\|\sum_{\tau=1}^{t-1}\phi(x_h^\tau, a_h^\tau)\cdot\big[V_{h+1}^t(x_{h+1}^\tau) - (\mathbb{P}_h V_{h+1}^t)(x_h^\tau, a_h^\tau)\big]\bigg\|^2_{(\Lambda_h^t)^{-1}} \\
&\quad\le 4H^2\cdot\Gamma_K(T,\lambda) + 11H^2 + 4H^2\cdot\log N_\infty(\epsilon^*; h+1, B_T) + 8H^2\cdot\log(TH).
\end{aligned} \tag{C.26}$$

Thus, combining (C.17), (C.20), (C.21), (C.23), and (C.26), we obtain that

$$\begin{aligned}
&\big|\phi(z)^\top(\widehat{\theta}_h^t - \overline{\theta}_h^t)\big| \\
&\quad\le H\cdot\big\{\big[4\cdot\Gamma_K(T,\lambda) + 4\cdot\log N_\infty(\epsilon^*; h+1, B_T) + 8\cdot\log(TH) + 11\big]^{1/2} + \sqrt{\lambda}R_Q\big\}\cdot b_h^t(z) \\
&\quad\le H\cdot\big[8\cdot\Gamma_K(T,\lambda) + 8\cdot\log N_\infty(\epsilon^*; h+1, B_T) + 16\cdot\log(TH) + 22 + 2R_Q^2\lambda\big]^{1/2}\cdot b_h^t(z) \\
&\quad\le B_T\cdot b_h^t(z) = \beta\cdot b_h^t(z)
\end{aligned} \tag{C.27}$$

holds uniformly for all $(t, h) \in [T] \times [H]$ with probability at least $1 - (2T^2H^2)^{-1}$, where the second inequality follows from the basic inequality $\sqrt{a} + \sqrt{b} \le \sqrt{2(a^2 + b^2)}$ and the last inequality follows from the assumption on B_T given in (4.5).

Finally, by (C.27) and the definition of the temporal-difference error δ_h^t in (5.1), we have

$$-\delta_h^t(z) = Q_h^t(z) - \phi(z)^\top\overline{\theta}_h^t \le \phi(z)^\top(\widehat{\theta}_h^t - \overline{\theta}_h^t) + \beta\cdot b_h^t(z) \le 2\beta\cdot b_h^t(z). \tag{C.28}$$

In addition, since $Q_{h+1}^t(z) \le H - h$ for all $z \in \mathcal{Z}$, we have $(\mathbb{T}_h^\star Q_{h+1}^t) \le H - h + 1$. Hence, we have

$$\begin{aligned}
\delta_h^t(z) &= \phi(z)^\top\overline{\theta}_h^t - \min\big\{\phi(z)^\top\widehat{\theta}_h^t + \beta\cdot b_h^t(z), H - h + 1\big\}^+ \\
&\le \max\big\{\phi(z)^\top\overline{\theta}_h^t - \phi(z)^\top\widehat{\theta}_h^t - \beta\cdot b_h^t(z), \phi(z)^\top\overline{\theta}_h^t - (H - h + 1)\big\} \le 0.
\end{aligned} \tag{C.29}$$

Therefore, combining (C.28) and (C.29), we conclude the proof of Lemma 5.2. $\square$

C.3 Proof of Lemma 5.3

Proof. Following Cai et al. (2019a), we prove this lemma by showing that $\{\zeta_{t,h}^1, \zeta_{t,h}^2\}_{(t,h)\in[T]\times[H]}$ can be written as a bounded Martingale difference sequence with respect to a filtration. In particular, we construct the filtration explicitly as follows. For any $(t,h) \in [T]\times[H]$, we define σ-algebras $\mathcal{F}_{t,h,1}$ and $\mathcal{F}_{t,h,2}$ respectively as

$$\begin{aligned}\mathcal{F}_{t,h,1} &= \sigma\Big(\{(x_i^\tau, a_i^\tau)\}_{(\tau,i)\in[t-1]\times[H]} \bigcup \{(x_i^t, a_i^t)\}_{i\in[h]}\Big),\\ \mathcal{F}_{t,h,2} &= \sigma\Big(\{(x_i^\tau, a_i^\tau)\}_{(\tau,i)\in[t-1]\times[H]} \bigcup \{(x_i^t, a_i^t)\}_{i\in[h]} \bigcup \{x_{h+1}^t\}\Big),\end{aligned} \tag{C.30}$$

where $\sigma(\cdot)$ denotes the σ-algebra generated by a finite set. Moreover, for any $t \in [T]$, $h \in [H]$ and $m \in [2]$, we define the timestep index $\tau(t,h,m)$ as

$$\tau(t,h,m) = (t-1)\cdot 2H + (h-1)\cdot 2 + m, \tag{C.31}$$

which offers an partial ordering over the triplets $(t,h,m) \in [T]\times[H]\times[2]$. Moreover, by the definitions in (C.30), for any (t,h,m) and (t',h',m') satisfying $\tau(k,h,m) \le \tau(k',h',m')$, it holds that $\mathcal{F}_{k,h,m} \subseteq \mathcal{F}_{k',h',m'}$. Thus, the sequence if σ-algebras $\{\mathcal{F}_{t,h,m}\}_{(t,h,m)\in[T]\times[H]\times[2]}$ forms a filtration.

Furthermore, for any $(t,h) \in [T]\times[H]$, since both Q_h^t and V_h^t are obtained based on the trajectories of the first $(t-1)$ episodes, they are both measurable with respect to $\mathcal{F}_{t,1,1}$, which is a subset of $\mathcal{F}_{t,h,m}$ for all $h \in [H]$ and $m \in [2]$. Thus, by (C.30), $\zeta_{t,h}^1$ defined in (5.2) and $\zeta_{t,h}^2$ defined in (5.3) are measurable with respect to $\mathcal{F}_{t,h,1}$ and $\mathcal{F}_{t,h,2}$, respectively. In addition, note that $a_h^t \sim \pi_h^t(\,|\,x_h^t)$ and that $x_{h+1}^t \sim \mathbb{P}_h(\cdot\,|\,x_h^t, a_h^t)$. Thus, we have

$$\mathbb{E}\big[\zeta_{t,h}^1 \,\big|\, \mathcal{F}_{t,h-1,2}\big] = 0, \qquad \mathbb{E}\big[\zeta_{t,h}^2 \,\big|\, \mathcal{F}_{t,h,1}\big] = 0, \tag{C.32}$$

where we identify $\mathcal{F}_{t,0,2}$ with $\mathcal{F}_{t-1,H,2}$ for all $t \ge 2$ and let $\mathcal{F}_{1,0,2}$ be the empty set. Combining (C.31) and (C.32), we can define a Martingale $\{M_{t,h,m}\}_{(t,h,m)\in[T]\times[H]\times[2]}$ indexed by $\tau(t,k,m)$ defined in (C.31) as follows. For any $(t,h,m) \in [T]\times[H]\times[2]$, we define

$$M_{t,h,m} = \bigg\{\sum_{(s,g,\ell)} \zeta_{s,g}^\ell \colon \tau(s,g,\ell) \le \tau(t,h,m)\bigg\}, \tag{C.33}$$

that is, $M_{t,h,m}$ is the sum of all terms of the form $\zeta_{s,g}^\ell$ defined in (5.2) or (5.3) such that its timestep index $\tau(s,g,\ell)$ is no greater than $\tau(t,h,m)$. By definition, it holds that

$$M_{K,H,2} = \sum_{t=1}^T \sum_{h=1}^H (\zeta_{t,h}^1 + \zeta_{t,h}^2). \tag{C.34}$$

Moreover, since V_h^t, Q_h^t, $V_h^{\pi^t}$, and $Q_h^{\pi^t}$ all takes values in $[0,H]$, we have $|\zeta_{t,h}^1| \le 2H$ and $|\zeta_{t,h}^2| \le 2H$ for all $(t,h) \in [T]\times[H]$. This means that the Martingale $M_{t,h,m}$ defined in (C.33) has uniformly bounded differences. Thus, applying the Azuma-Hoeffding inequality (Azuma, 1967) to $M_{T,H,2}$ in (C.34), we obtain that

$$\mathbb{P}\bigg(\bigg|\sum_{t=1}^T \sum_{h=1}^H (\zeta_{t,h}^1 + \zeta_{t,h}^2)\bigg| > t\bigg) \le 2\exp\bigg(\frac{-t^2}{16TH^3}\bigg) \tag{C.35}$$

holds for all $t > 0$. Finally, we set the right-hand side of (C.35) to ζ for some $\zeta \in (0,1)$, which yields $t = \sqrt{16TH^3 \cdot \log(2/\zeta)}$. Thus, we obtain that

$$\left| \sum_{t=1}^{T} \sum_{h=1}^{H} (\zeta_{t,h}^1 + \zeta_{t,h}^2) \right| \le \sqrt{16TH^3 \cdot \log(2/\zeta)}$$

with probability at least $1 - \zeta$, which concludes the proof. $\square$

C.4 Proof of Lemma 5.4

Proof. The proof of this lemma utilizes the connection between overparameterized neural network and NTK. Notice that we denote $z = (x, a)$ and $\mathcal{Z} = \mathcal{S} \times \mathcal{A}$. Also recall that $(b^{(0)}, W^{(0)})$ is the initial value of the network parameters obtained by the symmetric initialization scheme introduced in §2.3. Thus, $f(\cdot; W^{(0)})$ is a zero function. For any $(t,h) \in [T] \times [H]$, since $\widehat{W}_h^t$ is the global minimizer of loss function L_h^t defined in (3.9), we have

$$\begin{aligned} L_h^t(\widehat{W}_h^t) &= \sum_{\tau=1}^{t-1} \big[r_h(x_h^\tau, a_h^\tau) + V_{h+1}^t(x_{h+1}^\tau) - f(x_h^\tau, a_h^\tau; \widehat{W}_h^t) \big]^2 + \lambda \cdot \big\| \widehat{W}_h^t - W^{(0)} \big\|_2^2 \\ &\le L_h^t(W^{(0)}) = \sum_{\tau=1}^{t-1} \big[r_h(x_h^\tau, a_h^\tau) + V_{h+1}^t(x_{h+1}^\tau) \big]^2 \le (H-h+1)^2 \cdot (t-1) \le TH^2, \end{aligned} \tag{C.36}$$

where the second to last inequality follows from the facts that V_{h+1}^t is bounded by $H - h$ and that $r_h \in [0,1]$. Thus, (C.36) implies that

$$\big\| \widehat{W}_h^t - W^{(0)} \big\|_2^2 \le TH^2/\lambda, \qquad \forall (t,h) \in [T] \times [H]. \tag{C.37}$$

That is, each $\widehat{W}_h^t$ belongs to the Euclidean ball $\mathcal{B} = \{W \in \mathbb{R}^{2md} \colon \|W - W^{(0)}\|_2 \le H\sqrt{T/\lambda}\}$. Here the regularization parameter λ is does not depend on m and will be determined later. Notice that the radius of $\mathcal{B}$ does not depend on m. When m is sufficiently large, it can be shown that $f(\cdot, W)$ is close to its linearization $\widehat{f}(\cdot; W) = \langle \varphi(\cdot; W^{(0)}), W - W^{(0)} \rangle$ for all $W \in \mathcal{B}$, where $\varphi(\cdot; W) = \nabla_W f(\cdot; W)$.

Furthermore, recall that the temporal-difference error δ_h^t is defined as

$$\delta_h^t = r_h + \mathbb{P}_h V_{h+1}^t - Q_h^t = \mathbb{T}_h^\star Q_{h+1}^t - Q_h^t.$$

Under Assumption 4.5, we have $\mathbb{T}_h^\star Q_{h+1}^t \in \mathcal{Q}^\star$ for all $(t,h) \in [T] \times [H]$, where $\mathcal{Q}^\star$ is defined in (4.11). That is, for all $(t,h) \in [T] \times [H]$, there exists a function $\alpha_h^t \colon \mathbb{R}^d \to \mathbb{R}^d$ such that

$$(\mathbb{T}_h^\star Q_{h+1}^t)(z) = \int_{\mathbb{R}^d} \mathrm{act}'(w^\top z) \cdot z^\top \alpha_h^t(w) \, \mathrm{d}p_0(w), \qquad \forall (t,h) \in [T] \times [H], \forall z \in \mathcal{Z}. \tag{C.38}$$

Moreover, it holds that $\|\alpha_h^t\|_{2,\infty} = \sup_w \|\alpha_h^t(w)\|_2 \le R_Q H/\sqrt{d}$.

Now we are ready to bound the temporal-difference error δ_h^t defined in (5.1). Our proof is decomposed into three steps.

Step I. In the first step, we show that, with high probability, $\mathbb{T}_h^\star Q_{h+1}^t$ can be well-approximated by the class of linear functions of $\varphi(\cdot; W^{(0)})$ with respect to the ℓ_∞-norm.

Specifically, by Proposition C.1 in Gao et al. (2019), with probability at least $1-m^{-2}$ over the randomness of initialization, for any $(t,h)\in[T]\times[H]$, there exists a function $\widetilde{Q}_h^t\colon \mathcal{Z}\to\mathbb{R}$ that can be written as

$$\widetilde{Q}_h^t(z)=\frac{1}{\sqrt{m}}\sum_{j=1}^m \mathrm{act}'(\langle W_j^{(0)},z\rangle)\cdot z^\top\alpha_j, \tag{C.39}$$

where $\|\alpha_j\|_2\le R_Q/\sqrt{dm}$ for all $j\in[m]$ and $\{W_j^{(0)}\}_{j\in[2m]}$ are the random weights generated in the symmetric initialization scheme. Moreover, $\widetilde{Q}_h^t$ satisfies that

$$\|\widetilde{Q}_h^t-\mathbb{T}_h^\star Q_{h+1}^t\|_\infty\le 10C_{\mathrm{act}}R_QH\cdot\sqrt{\log(mTH)/m}. \tag{C.40}$$

Moreover, for any $j\in[2m]$, let $W_j^{(0)}$ and $b_j^{(0)}$ be the j-th component of $b^{(0)}$ and $W^{(0)}$, respectively.

Now we show that $\widetilde{Q}_h^t$ in (C.39) can be written as $\varphi(\cdot;W^{(0)})^\top(\widetilde{W}_h^t-W^{(0)})$ for some $\widetilde{W}_h^t\in\mathbb{R}^{2md}$. To this end, we define $\widetilde{W}_h^t=(\widetilde{W}_1,\ldots\widetilde{W}_{2m})\in\mathbb{R}^{2md}$ as follows. For any $j\in[m]$, we let $\widetilde{W}_j=W_j^{(0)}+b_j^{(0)}\cdot\alpha_j/\sqrt{2}$, and for any $j\in\{m+1,\ldots,2m\}$, we let $\widetilde{W}_j=W_j^{(0)}+b_j^{(0)}\cdot\alpha_{j-m}/\sqrt{2}$. Then, by the symmetric initialization scheme, we have

$$\begin{aligned}
\widetilde{Q}_h^t(z)&=\frac{1}{\sqrt{2m}}\sum_{j=1}^m\sqrt{2}\cdot(b_j^{(0)})^2\cdot\mathrm{act}'\big(\langle W_j^{(0)},z\rangle\big)\cdot z^\top\alpha_j\\
&=\frac{1}{\sqrt{2m}}\sum_{j=1}^m 1/\sqrt{2}\cdot(b_j^{(0)})^2\cdot\mathrm{act}'\big(\langle W_j^{(0)},z\rangle\big)\cdot z^\top\alpha_j+\frac{1}{\sqrt{2m}}\sum_{j=1}^m 1/\sqrt{2}\cdot(b_j^{(0)})^2\cdot\mathrm{act}'\big(\langle W_j^{(0)},z\rangle\big)\cdot z^\top\alpha_{j-m}\\
&=\frac{1}{\sqrt{2m}}\sum_{j=1}^{2m}b_j^{(0)}\cdot\mathrm{act}'\big(\langle W_j^{(0)},z\rangle\big)\cdot z^\top\big(\widetilde{W}_j-W_j^{(0)}\big)=\varphi(z;W^{(0)})^\top\big(\widetilde{W}_h^t-W^{(0)}\big).
\end{aligned}\tag{C.41}$$

Besides, since $\|\alpha_j\|_2\le R_QH/\sqrt{dm}$, we have $\|\widetilde{W}_h^t-W^{(0)}\|_2\le R_QH/\sqrt{d}$.

Therefore, for all $(t,h)\in[T]\times[H]$, we have constructed $\widetilde{Q}_h^t$ that is linear $\varphi(\cdot;W^{(0)})$. Moreover, with probability at least $1-m^{-2}$ over the randomness of initialization, $\widetilde{Q}_h^t$ is close to $\mathbb{T}_h^\star Q_{h+1}^t$ in the sense that (C.40) holds uniformly for all $(t,h)\in[T]\times[H]$. Thus, we conclude the first step.

Step II. In the second step, we show Q_h^t used in Algorithm 3 can be well-approximated using functions based on the feature mapping $\varphi(\cdot;W^{(0)})$.

Recall that the bonus in Q_h^t utilizes matrix Λ_h^t defined in (3.11), which involves the neural tangent features $\{\varphi(\cdot;\widehat{W}_h^\tau)\}_{\tau\in[T]}$. Similar to Λ_h^t, we define $\overline{\Lambda}_h^t$ as

$$\overline{\Lambda}_h^t=\lambda\cdot I_{2md}+\sum_{\tau=1}^{t-1}\varphi(x_h^\tau,a_h^\tau;W^{(0)})\varphi(x_h^\tau,a_h^\tau;W^{(0)})^\top, \tag{C.42}$$

which adopts the same feature mapping $\varphi(\cdot;W^{(0)})$. To simplify the notation, hereafter, we use $\varphi(\cdot)$ to denote $\varphi(\cdot;W^{(0)})$ when its meaning is clear from the text. Besides, for any $(t,h)\in[T]\times[H]$, we define the response vector $y_h^t\in\mathbb{R}^{t-1}$ by letting its entries be

$$[y_h^t]_\tau=r_h(x_h^\tau,a_h^\tau)+V_{h+1}^t(x_{h+1}^\tau),\qquad\forall\tau\in[t-1]. \tag{C.43}$$

We define the feature matrix $\Phi_h^t \in \mathbb{R}^{(t-1)\times 2md}$ by

$$\Phi_h^t = \big[\varphi(x_h^1, a_h^1)^\top, \ldots, \varphi(x_h^{t-1}, a_h^{t-1})^\top\big]^\top. \tag{C.44}$$

Hence, by (C.42) and (C.44), we have $\overline{\Lambda}_h^t = \lambda \cdot I_{2md} + (\Phi_h^t)^\top \Phi_h^t$. Similar to the bonus function b_h^t defined in (3.12), we define

$$\overline{b}_h^t = \big[\varphi(x,a)^\top (\overline{\Lambda}_h^t)^{-1} \varphi(x,a)\big]^{1/2} = \|\varphi(x,a)\|_{(\overline{\Lambda}_h^t)^{-1}}. \tag{C.45}$$

Similar to L_h^t defined in (3.9), we define another least-squares loss function $\overline{L}_h^t \colon \mathbb{R}^{2md} \to \mathbb{R}$ as

$$\overline{L}_h^t(W) = \sum_{\tau=1}^{t-1} \big[r_h(x_h^\tau, a_h^\tau) + V_{h+1}^t(x_{h+1}^\tau) - \big\langle \varphi(x_h^\tau, a_h^\tau), W - W^{(0)} \big\rangle\big]^2 + \lambda \cdot \|W - W^{(0)}\|_2^2 \tag{C.46}$$

and let $\overline{W}_h^t$ be its global minimizer. By direct computation, $\overline{W}_h^t$ can be written in closed form as

$$\overline{W}_h^t = W^{(0)} + (\overline{\Lambda}_h^t)^{-1} (\Phi_h^t)^\top y_h^t, \tag{C.47}$$

where $\overline{\Lambda}_h^t$, Φ_h^t, and y_h^t are defined respectively in (C.42), (C.44), and (C.43). Similar to (C.36), utilizing the fact that $\overline{L}_h^t(\overline{W}_h^t) \le \overline{L}_h^t(W^{(0)})$, we also have $\|\overline{W}_h^t - W^{(0)}\|_2 \le H\sqrt{T/\lambda}$. Then, similar to Q_h^t constructed in Algorithm 3, we combine $\overline{b}_h^t$ in (C.45) and $\overline{W}_h^t$ in (C.47) to define $\overline{Q}_h^t \colon \mathcal{Z} \to \mathbb{R}$ as

$$\overline{Q}_h^t(x,a) = \min\big\{\varphi(x,a)^\top (\overline{W}_h^t - W^{(0)}) + \beta \cdot \overline{b}_h^t(x,a), H - h + 1\big\}^+. \tag{C.48}$$

Note that $\overline{Q}_h^t$ share the same form as Q in (4.12). Thus, we have $\overline{Q}_h^t \in \mathcal{Q}_{\mathrm{ucb}}(h, H\sqrt{T/\lambda}, B)$ for any $B \ge \beta$. Moreover, we define $\overline{V}_h^t(\cdot) = \max_{a\in\mathcal{A}} \overline{Q}_h^t(\cdot, a)$.

In the following, we aim to show that $\overline{Q}_h^t$ is close to Q_h^t when m is sufficiently large. When this is true, $\overline{V}_h^t$ is also close to V_h^t. To bound $Q_h^t - \overline{Q}_h^t$, since the truncation operator is non-expansive, by triangle inequality, we have

$$\|Q_h^t - \overline{Q}_h^t\|_\infty \le \underbrace{\big\|f(\cdot; \widehat{W}_h^t) - \varphi(\cdot)^\top (\overline{W}_h^t - W^{(0)})\big\|_\infty}_{\text{(i)}} + \underbrace{\beta \cdot \|b_h^t - \overline{b}_h^t\|_\infty}_{\text{(ii)}}. \tag{C.49}$$

Recall that we define $\mathcal{B} = \{W \in \mathbb{R}^{2md} \colon \|W - W^{(0)}\|_2 \le H\sqrt{T/\lambda}\}$. To bound the two terms on the right-hand side of (C.49), we utilize the following lemma that quantifies the perturbation of $f(\cdot; W)$ and $\varphi(\cdot; W)$ within $W \in \mathcal{B}$.

Lemma C.2. When $TH^2 = \mathcal{O}(m \cdot \log^{-6} m)$, with probability at least $1 - m^{-2}$ with respect to the randomness of initialization, for any $W \in \mathcal{B}$ and any $z \in \mathcal{Z}$, we have

$$\big|f(z, W) - \varphi(z, W^{(0)})^\top (W - W^{(0)})\big| \le \overline{C} \cdot T^{2/3} \cdot H^{4/3} \cdot m^{-1/6} \cdot \sqrt{\log m},$$
$$\big\|\varphi(z, W) - \varphi(z, W^{(0)})\big\|_2 \le \overline{C} \cdot (TH^2/m)^{1/6} \cdot \sqrt{\log m}, \qquad \|\varphi(z, W)\|_2 \le \overline{C}.$$

Proof. See Allen-Zhu et al. (2018b); Gao et al. (2019); Cai et al. (2019b) for a detailed proof. More specifically, this lemma is obtained from Lemmas F.1 and F.2 in Cai et al. (2019b), which are further based on results in Allen-Zhu et al. (2018b); Gao et al. (2019). $\square$

By Lemma C.2 and triangle inequality, Term (i) on the right-hand side of (C.49) is bounded by

$$\begin{aligned}\text{Term (i)} &\le \big\|f(\cdot;\widehat{W}_h^t) - \varphi(\cdot)^\top(\widehat{W}_h^t - W^{(0)})\big\|_\infty + \big\|\varphi(\cdot)^\top(\widehat{W}_h^t - \overline{W}_h^t)\big\|_\infty \\ &\le \overline{C}\cdot T^{2/3}\cdot H^{4/3}\cdot m^{-1/6}\cdot\sqrt{\log m} + \overline{C}\cdot\big\|\widehat{W}_h^t - \overline{W}_h^t\big\|_2. \end{aligned} \tag{C.50}$$

To bound $\|\widehat{W}_h^t - \overline{W}_h^t\|_2$, notice that $\widehat{W}_h^t$ and $\overline{W}_h^t$ are the global minimizers of L_h^t in (3.9) and $\overline{L}_h^t$ in (C.46), respectively. Thus, by the first-order optimality condition, we have

$$\lambda\cdot\big(\widehat{W}_h^t - W^{(0)}\big) = \sum_{\tau=1}^{t-1}\big\{[y_h^t]_\tau - f(z_h^\tau;\widehat{W}_h^t)\big\}\cdot\varphi(z_h^\tau;\widehat{W}_h^t), \tag{C.51}$$

$$\lambda\cdot(\overline{W}_h^t - W^{(0)}) = \sum_{\tau=1}^{t-1}\big\{[y_h^t]_\tau - \big\langle\varphi(z_h^\tau;W^{(0)}),\overline{W}_h^t - W^{(0)}\big\rangle\big\}\cdot\varphi(z_h^\tau;W^{(0)}), \tag{C.52}$$

where $[y_h^t]_\tau$ is defined in (C.43) and $z_h^\tau = (x_h^\tau, a_h^\tau)$. In addition, by the definition of $\overline{\Lambda}_h^t$ in (C.42), (C.52) can be equivalently written as

$$\overline{\Lambda}_h^t\big(\overline{W}_h^t - W^{(0)}\big) = \sum_{\tau=1}^{t-1}[y_h^t]_\tau\cdot\varphi(z_h^\tau;W^{(0)}). \tag{C.53}$$

Similarly, for (C.51), by direct computation we have

$$\begin{aligned}\overline{\Lambda}_h^t\big(\widehat{W}_h^t - W^{(0)}\big) &= \sum_{\tau=1}^{t-1}[y_h^t]_\tau\cdot\varphi\ (z_h^\tau;\widehat{W}_h^t) \\ &\quad + \sum_{\tau=1}^{t-1}\Big[\big\langle\varphi(z_h^\tau;W^{(0)}),\widehat{W}_h^t - W^{(0)}\big\rangle\cdot\varphi(z_h^\tau;W^{(0)}) - f(z_h^\tau;\widehat{W}_h^t)\cdot\varphi(z_h^\tau;\widehat{W}_h^t)\Big].\end{aligned} \tag{C.54}$$

For any $\tau\in[t-1]$, we have

$$\begin{aligned}&\big\langle\varphi(z_h^\tau;W^{(0)}),\widehat{W}_h^t - W^{(0)}\big\rangle\cdot\varphi(z_h^\tau;W^{(0)}) - f\big(z_h^\tau;\widehat{W}_h^t\big)\cdot\varphi(z_h^\tau;\widehat{W}_h^t) \\ &\quad = \big\langle\varphi(z_h^\tau;W^{(0)}),\widehat{W}_h^t - W^{(0)}\big\rangle\cdot\big[\varphi(z_h^\tau;W^{(0)}) - \varphi(z_h^\tau;\widehat{W}_h^t)\big] \\ &\qquad + \big[\big\langle\varphi(z_h^\tau;W^{(0)}),\widehat{W}_h^t - W^{(0)}\big\rangle - f(z_h^\tau;\widehat{W}_h^t)\big]\cdot\varphi(z_h^\tau;\widehat{W}_h^t).\end{aligned} \tag{C.55}$$

Thus, applying Lemma C.2 to (C.55), we have

$$\begin{aligned}&\Big\|\big\langle\varphi(z_h^\tau;W^{(0)}),\widehat{W}_h^t - W^{(0)}\big\rangle\cdot\varphi(z_h^\tau;W^{(0)}) - f(z_h^\tau;\widehat{W}_h^t)\cdot\varphi(z_h^\tau;\widehat{W}_h^t)\Big\|_2 \\ &\quad\le \big\|\varphi(z_h^\tau;W^{(0)})\big\|_2\cdot\big\|\widehat{W}_h^t - W^{(0)}\big\|_2\cdot\big\|\varphi(z_h^\tau;W^{(0)}) - \varphi(z_h^\tau;\widehat{W}_h^t)\big\| \\ &\qquad + \big|\big\langle\varphi(z_h^\tau;W^{(0)}),\widehat{W}_h^t - W^{(0)}\big\rangle - f(z_h^\tau;\widehat{W}_h^t)\big|\cdot\big\|\varphi(z_h^\tau;\widehat{W}_h^t)\big\|_2 \\ &\quad\le 2\overline{C}^2\cdot T^{2/3}\cdot H^{4/3}\cdot m^{-1/6}\cdot\sqrt{\log m}\cdot\lambda^{-1/2},\end{aligned} \tag{C.56}$$

where we utilize the fact that $\|\widehat{W}_h^t - W^{(0)}\|_2 \le H\sqrt{T/\lambda} \le H\sqrt{T}$. Thus, combining (C.53), (C.54), and (C.56), we have

$$
\begin{aligned}
&\|\overline{\Lambda}_h^t(\widehat{W}_h^t - \overline{W}_h^t)\|_2 \\
&\quad \le \Big\| \sum_{\tau=1}^{t-1} [y_h^t]_\tau \cdot \big[\varphi(z_h^\tau; \widehat{W}_h^t) - \varphi(z_h^\tau; W^{(0)})\big] \Big\|_2 + 2\overline{C}^2 \cdot T^{5/3} \cdot H^{4/3} \cdot m^{-1/6} \cdot \sqrt{\log m} \\
&\quad \le \overline{C} \cdot T^{7/6} \cdot H^{4/3} \cdot m^{-1/6} \cdot \sqrt{\log m} + 2\overline{C}^2 \cdot T^{5/3} \cdot H^{4/3} \cdot m^{-1/6} \cdot \sqrt{\log m},
\end{aligned} \tag{C.57}
$$

where in the last inequality we utilize the fact that $[y_h^t]_\tau \in [0, H]$. When T is sufficiently large, the second term in (C.57) dominates. Since the eigenvalues of $(\overline{\Lambda}_h^t)^{-1}$ are all bounded by $1/\lambda$, we have

$$
\|\widehat{W}_h^t - \overline{W}_h^t\|_2 \le \|(\overline{\Lambda}_h^t)^{-1}\|_{\mathrm{op}} \cdot \|\overline{\Lambda}_h^t(\widehat{W}_h^t - \overline{W}_h^t)\|_2 \le 1/\lambda \cdot \|\overline{\Lambda}_h^t(\widehat{W}_h^t - \overline{W}_h^t)\|_2. \tag{C.58}
$$

In the sequel, we set λ as

$$
\lambda = \overline{C}^2 \cdot (1 + 1/T) \in \big[\overline{C}^2, 2\overline{C}^2\big]. \tag{C.59}
$$

Thus, combining (C.50), (C.57), (C.58), and (C.59), we have

$$
\text{Term (i)} \le 4 \cdot T^{5/3} \cdot H^{4/3} \cdot m^{-1/6} \cdot \sqrt{\log m} \tag{C.60}
$$

where we use the fact that $\overline{C}^2/\lambda \le 1$.

Furthermore, to bound Term (ii), by the definitions of b_h^t and $\overline{b}_h^t$, for any $z \in \mathcal{Z}$, we have

$$
\begin{aligned}
\big|b_h^t(z) - \overline{b}_h^t(z)\big| &= \Big| \sqrt{\varphi(z; \widehat{W}_h^t)^\top (\Lambda_h^t)^{-1} \varphi(z; \widehat{W}_h^t)} - \sqrt{\varphi(z; W^{(0)})^\top (\overline{\Lambda}_h^t)^{-1} \varphi(z; W^{(0)})} \Big| \\
&\le \sqrt{\big|\varphi(z; \widehat{W}_h^t)^\top (\Lambda_h^t)^{-1} \varphi(z; \widehat{W}_h^t) - \varphi(z; W^{(0)})^\top (\overline{\Lambda}_h^t)^{-1} \varphi(z; W^{(0)})\big|},
\end{aligned} \tag{C.61}
$$

where the inequality follows from the basic inequality $|\sqrt{x} - \sqrt{y}| \le \sqrt{|x-y|}$. By triangle inequality

$$
\begin{aligned}
&\Big|\varphi(z; \widehat{W}_h^t)^\top (\Lambda_h^t)^{-1} \varphi(z; \widehat{W}_h^t) - \varphi(z; W^{(0)})^\top (\overline{\Lambda}_h^t)^{-1} \varphi(z; W^{(0)})\Big| \\
&\quad \le \Big| \big[\varphi(z; \widehat{W}_h^t) - \varphi(z; W^{(0)})\big]^\top (\Lambda_h^t)^{-1} \varphi(z; \widehat{W}_h^t) \Big| + \Big| \varphi(z; W^{(0)})^\top \big[(\Lambda_h^t)^{-1} - (\overline{\Lambda}_h^t)^{-1}\big] \varphi(z; \widehat{W}_h^t) \Big| \\
&\qquad + \Big| \varphi(z; W^{(0)})^\top (\overline{\Lambda}_h^t)^{-1} \big[\varphi(z; \widehat{W}_h^t) - \varphi(z; W^{(0)})\big] \Big|.
\end{aligned} \tag{C.62}
$$

Combining Hölder's inequality and Lemma C.2, we bound the first term on the right-hand side of (C.62) by

$$
\begin{aligned}
&\Big| \big[\varphi(z; \widehat{W}_h^t) - \varphi(z; W^{(0)})\big]^\top (\Lambda_h^t)^{-1} \varphi(z; \widehat{W}_h^t) \Big| \qquad \text{(C.63)} \\
&\quad \le \big\|\varphi(z; \widehat{W}_h^t) - \varphi(z; W^{(0)})\big\|_2 \cdot \big\|(\Lambda_h^t)^{-1}\big\|_{\mathrm{op}} \cdot \big\|\varphi(z; \widehat{W}_h^t)\big\|_2 \le \overline{C}^2 \cdot T^{1/6} \cdot H^{1/3} \cdot m^{-1/6} \cdot \lambda^{-1} \cdot \sqrt{\log m},
\end{aligned}
$$

where $\|(\Lambda_h^t)^{-1}\|_{\mathrm{op}}$ is the matrix operator norm of $(\Lambda_h^t)^{-1}$, which is bounded by $1/\lambda$. Similarly, for the third term, we also have

$$
\Big| \varphi(z; W^{(0)})^\top (\overline{\Lambda}_h^t)^{-1} \big[\varphi(z; \widehat{W}_h^t) - \varphi(z; W^{(0)})\big] \Big| \le \overline{C}^2 \cdot T^{1/6} \cdot H^{1/3} \cdot m^{-1/6} \cdot \lambda^{-1} \cdot \sqrt{\log m}. \tag{C.64}
$$

For the second term, since both Λ_h^t and $\overline{\Lambda}_h^t$ are invertible, we have

$$\begin{aligned}\big\|(\Lambda_h^t)^{-1} - (\overline{\Lambda}_h^t)^{-1}\big\|_{\mathrm{op}} &= \big\|(\Lambda_h^t)^{-1}(\Lambda_h^t - \overline{\Lambda}_h^t)(\overline{\Lambda}_h^t)^{-1}\big\|_{\mathrm{op}} \\ &\le \|(\Lambda_h^t)^{-1}\|_{\mathrm{op}} \cdot \|(\overline{\Lambda}_h^t)^{-1}\|_{\mathrm{op}} \cdot \|\Lambda_h^t - \overline{\Lambda}_h^t\|_{\mathrm{op}} \le \lambda^{-2} \cdot \|\Lambda_h^t - \overline{\Lambda}_h^t\|_{\mathrm{fro}}. \end{aligned} \tag{C.65}$$

By direct computation, we have

$$\begin{aligned}\|\Lambda_h^t - \overline{\Lambda}_h^t\|_{\mathrm{fro}} &= \Big\|\sum_{\tau} \\ &\le \sum_{\tau=1}^{t-1} \Big\|\varphi(z_h^\tau; \widehat{W}_h^{\tau+1})\varphi(z_h^\tau; \widehat{W}_h^{\tau+1})^\top - \varphi(z_h^\tau; W^{(0)})\varphi(z_h^\tau; W^{(0)})^\top\Big\|_{\mathrm{fro}} \\ &\le \sum_{\tau=1}^{t-1} \Big\|\big[\varphi(z_h^\tau; \widehat{W}_h^{\tau+1}) - \varphi(z_h^\tau; W^{(0)})\big]\varphi(z_h^\tau; \widehat{W}_h^{\tau+1})^\top + \varphi(z_h^\tau; W^{(0)})\big[\varphi(z_h^\tau; \widehat{W}_h^{\tau+1}) - \varphi(z_h^\tau; W^{(0)})\big]^\top\Big\|_{\mathrm{fro}}\end{aligned}$$

Hence, by Lemma C.2 we can bound $\|\Lambda_h^t - \overline{\Lambda}_h^t\|_{\mathrm{fro}}$ by

$$\begin{aligned}\|\Lambda_h^t - \overline{\Lambda}_h^t\|_{\mathrm{fro}} &\le 2(t-1) \cdot \overline{C}^2 \cdot T^{1/6} \cdot H^{1/3} \cdot m^{-1/6} \cdot \sqrt{\log m} \\ &\le 2\overline{C}^2 \cdot T^{7/6} \cdot H^{1/3} \cdot m^{-1/6} \cdot \sqrt{\log m}. \end{aligned} \tag{C.66}$$

Hence, combining (C.65) and (C.66), the second term on the right-hand side of (C.62) can be bounded by

$$\begin{aligned}&\Big|\varphi(z; W^{(0)})^\top\big[(\Lambda_h^t)^{-1} - (\overline{\Lambda}_h^t)^{-1}\big]\varphi(z; \widehat{W}_h^t)\Big| \\ &\qquad \le \big\|\varphi(z; W^{(0)})\big\|_2 \cdot \big\|\varphi(z; \widehat{W}_h^t)\big\|_2 \cdot \big\|(\Lambda_h^t)^{-1} - (\overline{\Lambda}_h^t)^{-1}\big\|_{\mathrm{op}} \\ &\qquad \le 2\overline{C}^4 \cdot T^{7/6} \cdot H^{1/3} \cdot m^{-1/6} \cdot \lambda^{-2} \cdot \sqrt{\log m}. \end{aligned} \tag{C.67}$$

Notice that λ defined in (C.59) satisfies that $\lambda \ge \overline{C}^2$. Thus, combining (C.61)-(C.64), and (C.67), we have

$$\big|b_h^t(z) - \overline{b}_h^t(z)\big| \le 2 \cdot T^{7/12} \cdot H^{1/6} \cdot m^{-1/12} \cdot (\log m)^{1/4}, \qquad \forall (t,h) \in [T] \times [H], \tag{C.68}$$

which establishes the second inequality in (5.11). Finally, combining (C.49), (C.60), and (C.68), we conclude that

$$\|Q_h^t - \overline{Q}_h^t\|_\infty \le 4 \cdot T^{5/3} \cdot H^{4/3} \cdot m^{-1/6} \cdot \sqrt{\log m} + 2\beta \cdot T^{7/12} \cdot H^{1/6} \cdot m^{-1/12} \cdot (\log m)^{1/4}.$$

Note that $\beta > 1$. When $m = \Omega(\beta^{12} \cdot T^{13} \cdot H^{14} \cdot (\log m)^3)$, the second term in the above inequality is the dominating term. Thus, we have

$$\sup_{x \in \mathcal{S}} \big|V_h^t(x) - \overline{V}_h^t(x)\big| \le \|Q_h^t - \overline{Q}_h^t\|_\infty \le 4\beta \cdot T^{7/12} \cdot H^{1/6} \cdot m^{-1/12} \cdot (\log m)^{1/4}. \tag{C.69}$$

This concludes the second step.

Step III. In the last step, we establish optimism by comparing $\varphi(\cdot)^\top(\overline{W}_h^t - W^{(0)})$ and function $\widetilde{Q}_h^t$ defined in (C.39), where $\varphi(\cdot)$ denotes $\varphi(\cdot; W^{(0)})$. By the definition of $\overline{\Lambda}_h^t$ in (C.42), we have

$$\widetilde{W}_h^t - W^{(0)} = (\overline{\Lambda}_h^t)^{-1} \cdot \big[\lambda \cdot \big(\widetilde{W}_h^t - W^{(0)}\big) + (\Phi_h^t)^\top \Phi_h^t \big(\widetilde{W}_h^t - W^{(0)}\big)\big],$$

where $\widetilde{W}_h^t$ is given in (C.41). Hence, combining (C.47), we have

$$\overline{W}_h^t - \widetilde{W}_h^t = -\lambda \cdot (\overline{\Lambda}_h^t)^{-1}\big(\widetilde{W}_h^t - W^{(0)}\big) + (\overline{\Lambda}_h^t)^{-1}(\Phi_h^t)^\top\big[y_h^t - \Phi_h^t\big(\widetilde{W}_h^t - W^{(0)}\big)\big]. \tag{C.70}$$

Thus, for any $z \in \mathcal{Z}$, by (C.70) we have

$$\begin{aligned}
&\varphi(z)^\top\big(\overline{W}_h^t - \widetilde{W}_h^t\big) \\
&\quad = \underbrace{-\lambda \cdot \varphi(z)^\top (\overline{\Lambda}_h^t)^{-1} \cdot \big(\widetilde{W}_h^t - W^{(0)}\big)}_{\text{(iii)}} + \underbrace{\varphi(z)^\top (\overline{\Lambda}_h^t)^{-1}(\Phi_h^t)^\top\big[y_h^t - \Phi_h^t\big(\widetilde{W}_h^t - W^{(0)}\big)\big]}_{\text{(iv)}}.
\end{aligned} \tag{C.71}$$

For Term (iii) on the right-hand side of (C.71), by Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\big|\lambda \cdot \varphi(z)^\top (\overline{\Lambda}_h^t)^{-1} \cdot \big(\widetilde{W}_h^t - W^{(0)}\big)\big| &\le \lambda \cdot \big\|\widetilde{W}_h^t - W^{(0)}\big\|_2 \cdot \big\|(\overline{\Lambda}_h^t)^{-1}\varphi(z)\big\|_2 \\
&\le \lambda \cdot R_Q H/\sqrt{d} \cdot \sqrt{\varphi(z)^\top (\overline{\Lambda}_h^t)^{-1}(\overline{\Lambda}_h^t)^{-1}\varphi(z)} \le R_Q H \cdot \sqrt{\lambda/d} \cdot \overline{b}_h^t(z).
\end{aligned} \tag{C.72}$$

For Term (iv) in (C.71), recall that $\widetilde{Q}_h^t(z) = \varphi(z)^\top(\widetilde{W}_h^t - W^{(0)})$. To simplify the notation, let $q^\star \in \mathbb{R}^{t-1}$ denote the vector whose τ-th entry is $(\mathbb{T}_h^\star Q_{h+1}^t)(x_h^\tau, a_h^\tau)$ for any $\tau \in [t-1]$. Then, by (C.40), for any $\tau \in [t-1]$, the τ-th entry of $\Phi_h^t(\widetilde{W}_h^t - W^{(0)})$ satisfies

$$\begin{aligned}
\big|[\Phi_h^t\big(\widetilde{W}_h^t - W^{(0)}\big)]_\tau - [q^\star]_\tau\big| &= \Big|\big[\Phi_h^t\big(\widetilde{W}_h^t - W^{(0)}\big)\big]_\tau - (\mathbb{T}_h^\star Q_{h+1}^t)(x_h^\tau, a_h^\tau)\Big| \\
&\le 10 C_{\text{act}} \cdot R_Q H \cdot \sqrt{\log(mTH)/m}.
\end{aligned}$$

Besides, for any $\tau \in [t-1]$, the τ-th entry of $(y_h^t - q^\star)$ can be written as

$$\begin{aligned}
[y_h^t]_\tau - [q^\star]_\tau &= r_h(x_h^\tau, a_h^\tau) + V_{h+1}^t(x_{h+1}^\tau) - \varphi(x_h^\tau, a_h^\tau)^\top \overline{\theta}_h^t = r_h(x_h^\tau, a_h^\tau) + V_{h+1}^t(x_{h+1}^\tau) - (\mathbb{T}_h^\star Q_{h+1}^t)(x_h^\tau, a_h^\tau) \\
&= V_{h+1}^t(x_{h+1}^\tau) - (\mathbb{P}_h V_{h+1}^t)(x_h^\tau, a_h^\tau).
\end{aligned} \tag{C.73}$$

Then, by the triangle inequality and (C.73), we have

$$\begin{aligned}
&\Big|\varphi(z)^\top (\overline{\Lambda}_h^t)^{-1}(\Phi_h^t)^\top\big[y_h^t - \Phi_h^t\big(\widetilde{W}_h^t - W^{(0)}\big)\big]\Big| \\
&\quad \le \Big|\varphi(z)^\top (\overline{\Lambda}_h^t)^{-1}(\Phi_h^t)^\top\big[y_h^t - q^\star\big]\Big| + \Big|\varphi(z)^\top (\overline{\Lambda}_h^t)^{-1}(\Phi_h^t)^\top\big[q^\star - \Phi_h^t\big(\widetilde{W}_h^t - W^{(0)}\big)\big]\Big| \\
&\quad \le \|\varphi(z)\|_{(\overline{\Lambda}_h^t)^{-1}} \cdot \Bigg\|\sum_{\tau=1}^{t-1} \varphi(x_h^\tau, a_h^\tau) \cdot \big[V_{h+1}^t(x_{h+1}^\tau) - (\mathbb{P}_h V_{h+1}^t)(x_h^\tau, a_h^\tau)\big]\Bigg\|_{(\overline{\Lambda}_h^t)^{-1}} \\
&\qquad + 10 C_{\text{act}} \cdot R_Q H \cdot \sqrt{\log(mTH)/m} \cdot \|\varphi(z)\|_{(\overline{\Lambda}_h^t)^{-1}}.
\end{aligned} \tag{C.74}$$

Recall that we have shown in **Step (II)** that, with probability at least $1 - m^2$ with respect to the randomness of initialization, (C.69) holds for all $(t, h) \in [T] \times [H]$. To simplify the notation, we denote

$$\mathtt{Err} = 4\beta \cdot T^{7/12} \cdot H^{1/6} \cdot m^{-1/12} \cdot (\log m)^{1/4}. \tag{C.75}$$

Moreover, we define functions $\Delta V_1 = V^t_{h+1} - \overline{V}^t_{h+1}$ and $\Delta V_2 = \mathbb{P}_h(V^t_{h+1} - \overline{V}^t_{h+1})$. Then (C.69) implies that $\sup_{x\in S} |\Delta V_1(x)| \le \mathtt{Err}$ and $\sup_{z\in\mathcal{Z}} |\Delta V_2(z)| \le \mathtt{Err}$. By the basic inequality $(a+b)^2 \le 2a^2 + 2b^2$, we have

$$
\begin{aligned}
&\bigg\| \sum_{\tau=1}^{t-1} \varphi(x_h^\tau, a_h^\tau) \cdot \big[V^t_{h+1}(x^\tau_{h+1}) - (\mathbb{P}_h V^t_{h+1})(x_h^\tau, a_h^\tau)\big] \bigg\|^2_{(\overline{\Lambda}^t_h)^{-1}} \\
&\quad \le 2 \underbrace{\bigg\| \sum_{\tau=1}^{t-1} \varphi(x_h^\tau, a_h^\tau) \cdot \big[\overline{V}^t_{h+1}(x^\tau_{h+1}) - (\mathbb{P}_h \overline{V}^t_{h+1})(x_h^\tau, a_h^\tau)\big] \bigg\|^2_{(\overline{\Lambda}^t_h)^{-1}}}_{\text{(v)}} \\
&\qquad + 2 \bigg\| \sum_{\tau=1}^{t-1} \varphi(x_h^\tau, a_h^\tau) \cdot \big[\Delta V_1(x^\tau_{h+1}) - \Delta V_2(x_h^\tau, a_h^\tau)\big] \bigg\|^2_{(\overline{\Lambda}^t_h)^{-1}} \\
&\quad \le 2 \cdot \text{Term (v)} + 8 \cdot \mathtt{Err}^2 \cdot T^2,
\end{aligned} \tag{C.76}
$$

where the last inequality follows from the fact that

$$
\begin{aligned}
&\bigg\| \sum_{\tau=1}^{t-1} \varphi(x_h^\tau, a_h^\tau) \cdot \big[\Delta V_1(x^\tau_{h+1}) - \Delta V_2(x_h^\tau, a_h^\tau)\big] \bigg\|^2_{(\overline{\Lambda}^t_h)^{-1}} \le 4\mathtt{Err}^2 \cdot \bigg\| \sum_{\tau=1}^{t-1} \varphi(x_h^\tau, a_h^\tau) \bigg\|^2_{(\overline{\Lambda}^t_h)^{-1}} \\
&\quad \le 4 \cdot \mathtt{Err}^2 \cdot (t-1) \cdot \lambda^{-1} \cdot \sum_{\tau=1}^{t-1} \|\varphi(x_h^\tau, a_h^\tau)\|_2^2 \le 4 \cdot \mathtt{Err}^2 \cdot (t-1)^2 \cdot \overline{C}^2 \cdot \lambda^{-1} \le 4 \cdot \mathtt{Err}^2 \cdot T^2.
\end{aligned}
$$

Here the second to last inequality follows from Lemma C.2 and the definition of λ.

Recall that we define $\overline{b}^t_h(z) = \|\varphi(z)\|_{(\overline{\Lambda}^t_h)^{-1}}$. Combining (C.73), (C.74), and (C.77), we have

$$
\begin{aligned}
&\big| \varphi(z)^\top (\overline{\Lambda}^t_h)^{-1} (\Phi^t_h)^\top \big[y^t_h - \Phi^t_h(\widetilde{W}^t_h - W^{(0)})\big] \big| \\
&\quad \le \overline{b}^t_h(z) \cdot \big[10 C_{\text{act}} \cdot R_Q H \cdot \sqrt{\log(mTH)/m} + \sqrt{2 \cdot \text{Term (v)}} + 2\sqrt{2} \cdot \mathtt{Err} \cdot T\big] \\
&\quad \le \overline{b}^t_h(z) \cdot \big[R_Q H + \sqrt{2 \cdot \text{Term (v)}}\big],
\end{aligned} \tag{C.77}
$$

where we apply the basic inequality $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$. Here in the last inequality we let m to be sufficiently large such that

$$
10 C_{\text{act}} \cdot R_Q H \cdot \sqrt{\log(mTH)/m} + 2\sqrt{2} \cdot \mathtt{Err} \cdot T \le R_Q H.
$$

In the following, we aim to bound Term (v) in (C.77) by combining the concentration of the self-normalized stochastic process and uniform concentration. To characterize the function class that contains each $\overline{V}^t_h$, we define $\widetilde{\varphi}\colon \mathcal{Z} \to \mathbb{R}$ by $\widetilde{\varphi}(z) = \varphi(z)/\overline{C}$. Then, conditioning on the event where the statements in Lemma C.2 are true, we have $\|\widetilde{\varphi}(z)\|_2 \le 1$ for all $z \in \mathcal{Z}$. Furthermore, we define a kernel function $\widetilde{K}\colon \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ by letting $\widetilde{K}(z, z') = \widetilde{\varphi}(z)^\top \widetilde{\varphi}(z')$ for all $z, z' \in \mathcal{Z}$. That is, $\widetilde{K}$ is the normalized version of the empirical NTK K_m. By construction, $\widetilde{K}$ is a bounded kernel such that $\sup_{z\in\mathcal{Z}} \widetilde{K}(z, z) \le 1$. We can also consider the maximal information gain in (4.2) for $\widetilde{K}$ and K_m. These two quantities are linked via

$$
\Gamma_{\widetilde{K}}(T, \sigma) = \Gamma_{K_m}\big(T, \overline{C}^2 \sigma\big), \qquad \forall \sigma > 0. \tag{C.78}
$$

Furthermore, we define $\widetilde{\lambda} = \lambda/\overline{C}^2$ and $\widetilde{\Lambda}_h^t = \overline{\Lambda}_h^t//\overline{C}^2$ for all $(t,h) \in [T] \times [H]$. By the definition of λ in (C.59), we have $\widetilde{\lambda} = 1 + 1/T \in [1,2]$. Moreover, by (C.42) we have

$$\widetilde{\Lambda}_h^t = \widetilde{\lambda} \cdot I_{2md} + \sum_{\tau=1}^{t-1} \widetilde{\varphi}(x_h^\tau, a_h^\tau)\widetilde{\varphi}(x_h^\tau, a_h^\tau)^\top.$$

Since $\widetilde{\lambda} > 1$, $\widetilde{\Lambda}_h^t$ is an invertible matrix and the eigenvalues of $(\widetilde{\Lambda}_h^t)^{-1}$ are all bounded above by one.

Using $\widetilde{\varphi}$ and $\widetilde{\Lambda}_h^t$, we rewrite each $\overline{Q}_h^t$ as follows. For $\overline{W}_h^t$ defined in (C.47), we have

$$\varphi(x,a)^\top \big(\overline{W}_h^t - W^{(0)}\big) = \overline{C} \cdot \widetilde{\varphi}(x,a)^\top \big(\overline{W}_h^t - W^{(0)}\big), \tag{C.79}$$

where $\overline{C} \cdot \|\overline{W}_h^t - W^{(0)}\|_2 \le \overline{C} \cdot H\sqrt{T/\lambda} \le H\sqrt{T}$ since $\lambda \ge (\overline{C})^2$. Meanwhile, we also have

$$\overline{b}_h^t(z) = \|\varphi(z)\|_{(\overline{\Lambda}_h^t)^{-1}} = \big[\widetilde{\varphi}(z)^\top \big(\widetilde{\Lambda}_h^t\big)^{-1} \widetilde{\varphi}(z)\big]^{1/2}. \tag{C.80}$$

Thus, combining (C.79) and (C.80), $\overline{Q}_h^t$ defined in (C.48) can be equivalently written as

$$\overline{Q}_h^t(z) = \min\Big\{\widetilde{\varphi}(z)^\top \overline{\vartheta}_h^t + \beta \cdot \big\|\widetilde{\varphi}(z)\big\|_{(\widetilde{\Lambda}_h^t)^{-1}}, H - h + 1\Big\}^+,$$

where $\overline{\theta}_h^t = \overline{C} \cdot (\overline{W}_h^t - W^{(0)})$, which satisfies $\|\overline{\vartheta}_h^t\|_2 \le H\sqrt{T}$.

Let $\mathcal{D}$ be a finite subset of $\mathcal{Z}$ with no more than T elements. For any fixed $\mathcal{D}$, we define

$$\widetilde{\Lambda}_{\mathcal{D}} = \widetilde{\lambda} \cdot I_{2dm} + \sum_{z \in \mathcal{D}} \widetilde{\varphi}(z)\widetilde{\varphi}(z)^\top \in \mathbb{R}^{2md \times 2md}. \tag{C.81}$$

For any $h \in [H]$, $R, B > 0$, we let $\widetilde{Q}_{\mathrm{ucb}}(h, R, B)$ consists of functions that take the form of

$$Q(\cdot) = \min\Big\{\widetilde{\varphi}(\cdot)^\top \vartheta + \beta \cdot \big\|\widetilde{\varphi}(\cdot)\big\|_{(\widetilde{\Lambda}_{\mathcal{D}})^{-1}}; H - h + 1\Big\}^+$$

for some $\vartheta \in \mathbb{R}^{2md}$ with $\|\vartheta\|_2 \le R$ and some $\mathcal{D} \subseteq \mathcal{Z}$. Then $\widetilde{Q}_{\mathrm{ucb}}(h, R, B)$ corresponds to the function class in (4.4) with the kernel being $\widetilde{K}$. Besides, we define $\widetilde{\mathcal{V}}_{\mathrm{ucb}}(h, R, B)$ as

$$\widetilde{\mathcal{V}}_{\mathrm{ucb}}(h, R, B) = \big\{V \colon V(\cdot) = \max_a Q(\cdot, a) \ \text{ for some } \ Q \in \widetilde{\mathcal{Q}}_{\mathrm{ucb}}(h, R, B)\big\}.$$

By definition, for all $h \in [H]$ and any $R, B > 0$, it holds that $\mathcal{Q}_{\mathrm{ucb}}(h, R, B) = \widetilde{\mathcal{Q}}_{\mathrm{ucb}}(h, \overline{C}R, B)$. Meanwhile, since $(\overline{C})^2 \le \lambda \le 2(\overline{C})^2$, for all $R > 0$, it holds that

$$\mathcal{Q}_{\mathrm{ucb}}(h, R, B) \subseteq \widetilde{\mathcal{Q}}_{\mathrm{ucb}}(h, R\sqrt{\lambda}, B) \subseteq \mathcal{Q}_{\mathrm{ucb}}(h, \sqrt{2}R, B). \tag{C.82}$$

Recall that we define $R_T = H\sqrt{2T/\lambda}$ and let $N_\infty(\epsilon; h, B)$ denote the ϵ-covering number of $\mathcal{Q}_{\mathrm{ucb}}(h, R_T, B)$ with respect to the ℓ_∞-norm on $\mathcal{Z}$. Moreover, hereafter, we denote $\epsilon^* = H/T$ and set $B = B_T$ which satisfy (4.13). Since we set $\beta = B_T$ in Algorithm 3, it holds for all $(t,h) \in [T] \times [H]$ that

$$\overline{Q}_h^t \in \widetilde{\mathcal{Q}}_{\mathrm{ucb}}(h, H\sqrt{T}, B) \subseteq \mathcal{Q}_{\mathrm{ucb}}(h, R_T, B), \qquad \overline{V}_h^t \in \widetilde{\mathcal{V}}_{\mathrm{ucb}}(h, H\sqrt{T}, B).$$

Now, to bound Term (v) in (C.77), similar to the analysis the proof of Lemma 5.2, we apply the concentration of self-normalized stochastic process and uniform concentration over $\widetilde{\mathcal{V}}_{\mathrm{ucb}}(h, H\sqrt{T}, B_T)$. Specifically, similar to (C.25) and (C.26), with probability at least $1-(2T^2H^2)^{-1}$, we have

$$
\begin{aligned}
\text{Term (v)} &= \Big\| \sum_{\tau=1}^{t-1} \varphi(x_h^\tau, a_h^\tau) \cdot \big[\overline{V}_{h+1}^t(x_{h+1}^\tau) - (\mathbb{P}_h \overline{V}_{h+1}^t)(x_h^\tau, a_h^\tau)\big] \Big\|_{(\overline{\Lambda}_h^t)^{-1}}^2 \\
&= \Big\| \sum_{\tau=1}^{t-1} \widetilde{\varphi}(x_h^\tau, a_h^\tau) \cdot \big[\overline{V}_{h+1}^t(x_{h+1}^\tau) - (\mathbb{P}_h \overline{V}_{h+1}^t)(x_h^\tau, a_h^\tau)\big] \Big\|_{(\widetilde{\Lambda}_h^t)^{-1}}^2 \\
&\le 4H^2 \cdot \Gamma_{\widetilde{K}}(T, \widetilde{\lambda}) + 11H^2 + 4H^2 \cdot \log N_\infty(\epsilon^*, h+1, B_T) + 8H^2 \cdot \log(TH). \qquad \text{(C.83)}
\end{aligned}
$$

Thus, combining (C.71), (C.72), (C.77), and (C.83), we obtain that

$$
\begin{aligned}
&\big|\varphi(z)^\top (\overline{W}_h^t - \widetilde{W}_h^t)\big| \\
&\quad \le \big|\text{Term (iii)}\big| + \big|\text{Term (iv)}\big| \le \big[R_Q H + \sqrt{2 \cdot \text{Term (v)}} + R_Q H \cdot \sqrt{\lambda/d}\big] \cdot \overline{b}_h^t(z) \\
&\quad \le H \cdot \Big\{ \big[8\Gamma_{K_m}(T, \lambda) + 22 + 8 \cdot \log N_\infty(\epsilon^*, h+1, B_T) + 16 \cdot \log(TH)\big]^{1/2} + R_Q \cdot (1 + \sqrt{\lambda/d}) \Big\} \cdot \overline{b}_h^t(z).
\end{aligned}
$$

Using the basic inequality $a + b \le \sqrt{2(a^2+b^2)}$, we have

$$
\begin{aligned}
&\big|\varphi(z)^\top (\overline{W}_h^t - \widetilde{W}_h^t)\big| \\
&\quad \le H \cdot \big[16\Gamma_{K_m}(T, \lambda) + 16 \cdot \log N_\infty(\epsilon^*, h+1, B_T) + 32 \cdot \log(TH) + 2R_Q^2 \cdot (1 + \sqrt{\lambda/d})^2\big]^{1/2} \cdot \overline{b}_h^t(z) \\
&\quad \le H \cdot \big[16\Gamma_{K_m}(T, \lambda) + 16 \cdot \log N_\infty(\epsilon^*, h+1, B_T) + 32 \cdot \log(TH) + 4R_Q^2 \cdot (1 + \lambda/d)\big]^{1/2} \cdot \overline{b}_h^t(z)
\end{aligned}
$$

By the choice of B_T in (4.13), we have

$$
\big|\varphi(z)^\top (\overline{W}_h^t - \widetilde{W}_h^t)\big| = \big|\varphi(z)^\top (\overline{W}_h^t - W^{(0)}) - \widetilde{Q}_h^t(z)\big| \le \beta \cdot \overline{b}_h^t(z)
$$

holds simultaneously for all $(t,h) \in [T] \times [H]$ and $z \in \mathcal{Z}$ with probability at least $1 - (2T^2H^2)^{-1}$.

Thus, combining this with (C.39) and (C.40), we have

$$
\big|\varphi(z)^\top (\overline{W}_h^t - W^{(0)}) - \mathbb{T}_h^\star Q_{h+1}^t(z)\big| \le \beta \cdot \overline{b}_h^t(z) + 10 C_{\mathrm{act}} \cdot R_Q H \cdot \sqrt{\log(mTH)/m}. \qquad \text{(C.84)}
$$

By the definition of $\overline{Q}_h^t$ in (C.48), we have

$$
\begin{aligned}
\overline{Q}_h^t(z) - \mathbb{T}_h^\star Q_{h+1}^t(z) &\le \varphi(z)^\top (\overline{W}_h^t - W^{(0)}) - \mathbb{T}_h^\star Q_{h+1}^t(z) + \beta \cdot \overline{b}_h^t(z) \\
&\le 2\beta \cdot \overline{b}_h^t(z) + 10 C_{\mathrm{act}} \cdot R_Q \cdot \sqrt{\log(mTH)/m}. \qquad \text{(C.85)}
\end{aligned}
$$

Moreover, since $\mathbb{T}_h^\star Q_{h+1}^t \le H - h + 1$, by (C.84) we have

$$
\begin{aligned}
\mathbb{T}_h^\star Q_{h+1}^t(z) - \overline{Q}_h^t(z) &= \mathbb{T}_h^\star Q_{h+1}^t(z) - \min\big\{\varphi(x,a)^\top \big(\overline{W}_h^t - W^{(0)}\big) + \beta \cdot \overline{b}_h^t(x,a), H - h + 1\big\}^+ \\
&= \max\big\{\mathbb{T}_h^\star Q_{h+1}^t(z) - \varphi(z)^\top \big(\overline{W}_h^t - W^{(0)}\big) - \beta \cdot \overline{b}_h^t(z), 0\big\}^+ \\
&\le 10 C_{\mathrm{act}} \cdot R_Q \cdot \sqrt{\log(mTH)/m}. \qquad \text{(C.86)}
\end{aligned}
$$

Let ι denote $T^{7/12} \cdot H^{1/12} \cdot m^{-1/12} \cdot (\log m)^{1/4}$. When m is sufficiently large, it holds that

$$10 C_{\text{act}} \cdot R_Q \cdot \sqrt{\log(mTH)/m} \le \iota \le \beta \cdot \iota.$$

Meanwhile, combining the definition of the TD error δ_h^t in (5.1) and (C.69), we have

$$\begin{aligned} \big|\delta_h^t(z) - \big[\mathbb{T}_h^\star Q_{h+1}^t(z) - \overline{Q}_h^t(z)\big]\big| &= \big|Q_h^t(z) - \overline{Q}_h^t(z)\big| \\ &\le 4\beta \cdot T^{7/12} \cdot H^{1/12} \cdot m^{-1/12} \cdot (\log m)^{1/4}. \end{aligned} \tag{C.87}$$

Finally, combining (C.85), (C.86), and (C.87), we establish that, with probability at least

$$\begin{aligned} &\delta_h^t(z) \le [\mathbb{T}_h^\star Q_{h+1}^t(z) - \overline{Q}_h^t(z)] + 4\beta \cdot \iota \le 5\beta \cdot \iota \\ &\delta_h^t(z) \ge [\mathbb{T}_h^\star Q_{h+1}^t(z) - \overline{Q}_h^t(z)] - 4\beta \cdot \iota \ge -2\beta \cdot \overline{b}_h^t(z) - 5\beta \cdot \iota \end{aligned}$$

hold for all $(t, h) \in [T] \times [H]$ simultaneously. Finally, combining this with (C.68), we have

$$-2\beta \cdot b_h^t - 9\beta \cdot \iota \le -2\beta \cdot \overline{b}_h^t - 5\beta \cdot \iota \le \delta_h^t(z) \le 5\beta \cdot \iota,$$

which, together with (C.68), concludes the proof of Lemma 5.4. $\square$

D Covering Number and Effective Dimension

In this section, we present the results on the covering number of the class of value functions and the effective dimension of the RKHS. Both of these results play a key role in establishing the regret bounds.

D.1 Covering Number of the Classes of Value Functions

For any $R, B > 0$, any $h \in [H]$, and fixed $\mathcal{D}$, we define $\mathcal{Q}_{\text{ucb}}(h, R, B)$ as the function class that contains functions on $\mathcal{Z}$ that take the form of

$$Q(z) = \min\Big\{\theta(z) + \beta \cdot \lambda^{-1/2}\big[K(z,z) - k_t(z)^\top (K_t + \lambda I)^{-1} k_t(z)\big]^{1/2}, H - h + 1\Big\}^+, \tag{D.1}$$

where $\theta \in \mathcal{H}$ satisfies $\|\theta\|_{\mathcal{H}} \le R$, $\beta \in [0, B]$, $h \in [H]$, and $\mathcal{D} = \{z^\tau = (x^\tau, a^\tau),\}_{\tau \in [t]}$ is a finite subset of $\mathcal{Z}$ with t elements, where $t \le T$. Here T is the total number of the episodes. Moreover, $K_t \in \mathbb{R}^{t\times t}$ and $k_t \colon \mathcal{Z} \to \mathbb{R}^t$ are defined similarly as in (3.7) based on state-action pairs in $\mathcal{D}$, that is,

$$K_t = [K(z^\tau, z^{\tau'})]_{\tau, \tau' \in [t]} \in \mathbb{R}^{t \times t}, \qquad k_t(z) = \big[K(z^1, z), \ldots K(z^t, z)\big]^\top \in \mathbb{R}^t.$$

By definition, Q in (D.1) is determined by $Q_0 \in \mathcal{H}$ and a bonus term constructed using $\mathcal{D}$. Thus, function Q_h^t constructed in Algorithm 2 belongs to $\mathcal{Q}_{\text{ucb}}(h, R, B)$ when $\beta \le B$ and $\|\alpha_h^t\|_{\mathcal{H}} \le R$. In the following, for any $\epsilon \in (0, 1)$, we let $\mathcal{C}(\mathcal{Q}_{\text{ucb}}(h, R, B), \epsilon)$ be the minimal ϵ-cover of $\mathcal{Q}_{\text{ucb}}(h, R, B)$ with respect to the ℓ_∞-norm on $\mathcal{Z}$. That is, for any $Q \in \mathcal{Q}_{\text{ucb}}(h, R, B)$, there exists $Q' \in \mathcal{C}(\mathcal{Q}_{\text{ucb}}(h, R, B), \epsilon)$ satisfying $\|Q - Q'\|_\infty \le \epsilon$. Moreover, among all function classes that possess such a property, $\mathcal{C}(\mathcal{Q}_{\text{ucb}}(h,R,B),\epsilon)$ has the smallest cardinality. Thus, by definition, $|\mathcal{C}(\mathcal{Q}_{\text{ucb}}(h,R,B),\epsilon)|$ is the ϵ-covering number of $\mathcal{Q}_{\text{ucb}}(h,R,B)$ with respect to the ℓ_∞-norm on $\mathcal{Z}$.

In addition, based on $\mathcal{Q}_{\text{ucb}}(h,R,B)$, we define function class $\mathcal{V}_{\text{ucb}}(h,R,B)$ as

$$\mathcal{V}_{\text{ucb}}(h,R,B) = \Big\{V\colon V(\cdot) = \max_a Q(\cdot,a) \ \text{ for some } \ Q \in \mathcal{Q}_{\text{ucb}}(h,R,B)\Big\}. \tag{D.2}$$

For any two value functions $V_1, V_2\colon \mathcal{S} \to \mathbb{R}$, we denote their supremum norm distance as

$$\text{dist}(V_1,V_2) = \sup_{x\in\mathcal{S}}\big|V_1(x) - V_2(x)\big|. \tag{D.3}$$

For any $\epsilon \in (0,1)$, we let $\mathcal{C}(\mathcal{V}_{\text{ucb}}(h,R,B),\epsilon)$ denote the minimal ϵ-cover of $\mathcal{V}_{\text{ucb}}(h,R,B)$ with respect to $\text{dist}(\cdot,\cdot)$ defined in (D.3).

The main result of this section is upper bounds on the size of $\mathcal{C}(\mathcal{V}_{\text{ucb}}(h,R,B),\epsilon)$ under the three eigenvalue decay conditions specified in Assumption 4.3, which are presented as follows.

Lemma D.1. Let Assumption 4.3 hold and λ be bounded in $[c_1,c_2]$, where both c_1 and c_2 are absolute constants. Then, for any $h \in [H]$, any $R, B > 0$, and any $\epsilon \in (0,1/e)$, there exists a positive constant C_N such that

$$\begin{aligned}
&\log\big|\mathcal{C}\big(\mathcal{V}_{\text{ucb}}(h,R,B),\epsilon\big)\big| \le \log\big|\mathcal{C}\big(\mathcal{Q}_{\text{ucb}}(h,R,B),\epsilon\big)\big| \qquad\qquad (\text{D.4})\\
&\quad\le \begin{cases} C_N\cdot\gamma\cdot\big[1+\log(R/\epsilon)\big] + C_N\cdot\gamma^2\cdot\big[1+\log(B/\epsilon)\big] & \text{case (i)},\\ C_N\cdot\big[1+\log(R/\epsilon)\big]^{1+1/\gamma} + C_N\cdot\big[1+\log(B/\epsilon)\big]^{1+2/\gamma} & \text{case (ii)},\\ C_N\cdot(R/\epsilon)^{2/[\gamma\cdot(1-2\tau)-1]}\cdot[1+\log(R/\epsilon)] + C_N\cdot(B/\epsilon)^{4/[\gamma\cdot(1-2\tau)-1]}\cdot[1+\log(B/\epsilon)] & \text{case (iii)},\end{cases}
\end{aligned}$$

where cases (i)–(iii) above correspond to the three eigenvalue decay conditions specified in Assumption 4.3, respectively. Moreover, here C_N in (D.4) is independent of T, H, R, and B, and only depends on C_ψ, C_1, C_2, c_1, c_2, γ, and τ.

Proof. For any fixed subset $\mathcal{D} = \{z^\tau\}_{\tau\in[t]}$ of $\mathcal{Z}$ with size $t\in[T]$, we define $\Phi_\mathcal{D}\colon\mathcal{H}\to\mathbb{R}^t$ and $\Lambda_\mathcal{D}\colon\mathcal{H}\to\mathcal{H}$ respectively as

$$\begin{aligned}
\Phi_\mathcal{D} &= \big[\phi(z^1)^\top,\phi(z^2)^\top\ldots,\phi(z^t)^\top\big]^\top,\\
\Lambda_\mathcal{D} &= \sum_{\tau=1}^t \phi(z^\tau)\phi(z^\tau)^\top + \lambda\cdot I_\mathcal{H} = \lambda\cdot I_\mathcal{H} + (\Phi_\mathcal{D})^\top\Phi_\mathcal{D},
\end{aligned} \tag{D.5}$$

where $\phi\colon\mathcal{Z}\to\mathcal{H}$ is the feature mapping of $\mathcal{H}$ and $\mathcal{I}_\mathcal{H}$ is the identity mapping on $\mathcal{H}$. Then, we can equivalently write $Q_1\in\mathcal{Q}_{\text{ucb}}(h,R,B)$ as

$$Q_1(z) = \phi(z)^\top\theta_1 + \beta\cdot\sqrt{\phi(z)^\top\Lambda_{\mathcal{D}_1}^{-1}\phi(z)}, \tag{D.6}$$

where $\theta_1\in\mathcal{H}$ has RKHS norms bounded by R, $\beta_1\in[0,B]$, and $\mathcal{D}_1$ is a finite subset of $\mathcal{Z}$ with size $t_1\le T$. Let $V_1(\cdot) = \max_{a\in\mathcal{A}} Q_1(\cdot,a)$. Combining (D.2) and (D.6), we can write $V_1\in\mathcal{V}_{\text{ucb}}(h,R,B)$ as

$$V_1(\cdot) = \min\Big\{\max_a\Big\{\phi(\cdot,a)^\top\theta_1 + \beta_1\cdot\sqrt{\phi(\cdot,a)^\top\Lambda_{\mathcal{D}_1}^{-1}\phi(\cdot,a)}\Big\}, H-h+1\Big\}^+, \tag{D.7}$$

Similar to V_1 in (D.7), consider any $V_2\colon \mathcal{S}\to\mathbb{R}$ that can be written as

$$V_2(\cdot) = \min\Big\{\max_a\Big\{f_1(\cdot,a)+\beta_2\cdot f_2(\cdot,a)\Big\}, H-h+1\Big\}^+, \tag{D.8}$$

where $Q_2 = f_1+\beta_2\cdot f_2$ for some $f_1, f_2\colon \mathcal{Z}\to\mathbb{R}$ and $\beta_2\in[0,B]$. Since both $\min\{\cdot, H-h+1\}^+$ and $\max_a$ are contractive mappings, by (D.7) and (D.8) we have

$$\operatorname{dist}(V_1,V_2)\le \sup_{(x,a)\in\mathcal{Z}}\left|\Big[\phi(x,a)^\top\theta_1+\beta_1\cdot\sqrt{\phi(x,a)^\top\Lambda_{\mathcal{D}_1}^{-1}\phi(x,a)}\Big]-\Big[f(x,a)+\beta_2\cdot f_2(x,a)\Big]\right| = \|Q_1-Q_2\|_\infty,$$

which implies that

$$\log\big|\mathcal{C}\big(\mathcal{V}_{\mathrm{ucb}}(h,R,B),\epsilon\big)\big|\le\log\big|\mathcal{C}\big(\mathcal{Q}_{\mathrm{ucb}}(h,R,B),\epsilon\big)\big|.$$

Moreover, by triangle inequality, we have

$$\begin{aligned}\|Q_1-Q_2\|_\infty \le \sup_{(x,a)\in\mathcal{Z}}\big|\phi(x,a)^\top\theta_1-f_2(x,a)\big|+|\beta_1-\beta_2|\cdot\sup_{(x,a)\in\mathcal{Z}}\big\|\phi(x,a)\big\|_{\Lambda_{\mathcal{D}_1}^{-1}}\\ +B\cdot\sup_{(x,a)\in\mathcal{Z}}\Big|\big\|\phi(x,a)\big\|_{\Lambda_{\mathcal{D}_1}^{-1}}-f_2(x,a)\Big|,\end{aligned} \tag{D.9}$$

where we denote $\|\phi(x,a)\|^2_{\Lambda_{\mathcal{D}_1}^{-1}} = \phi(x,a)^\top\Lambda_{\mathcal{D}_1}^{-1}\phi(x,a)$. Notice that by the reproducing property we have $\phi(x,a)^\top\theta = \langle\theta,\phi(x,a)\rangle_{\mathcal{H}} = \theta(x,a)$ for all $\theta\in\mathcal{H}$ and $(x,a)\in\mathcal{Z}$. Also note that

$$\big\|\phi(x,a)\big\|^2_{\Lambda_{\mathcal{D}_1}^{-1}}\le 1/\lambda\cdot\|\phi(x,a)\|^2\le 1/\lambda\cdot K(z,z)\le 1/\lambda.$$

Thus, by (D.9) we have

$$\begin{aligned}\|Q_1-Q_2\|_\infty\le \sup_{(x,a)\in\mathcal{Z}}\big|\theta_1(x,a)-f_1(x,a)\big|+|\beta_1-\beta_2|/\lambda\\ +B\cdot\sup_{(x,a)\in\mathcal{Z}}\Big|\big\|\phi(x,a)\big\|_{\Lambda_{\mathcal{D}_1}^{-1}}-f_2(x,a)\Big|.\end{aligned} \tag{D.10}$$

Thus, by (D.10), to get the covering number of $\mathcal{Q}_{\mathrm{ucb}}(h,R,B)$ with respect to $\operatorname{dist}(\cdot,\cdot)$, it suffices to bound the covering numbers of the RKHS norm ball $\{f\in\mathcal{H}\colon \|f\|_{\mathcal{H}}\le R\}$, interval $[0,B]$, and the set of functions that are of the form of $\|\phi(\cdot)\|_{\Lambda_{\mathcal{D}}^{-1}}$, respectively.

Notice that, by the definition in (D.5), $\Lambda_{\mathcal{D}}\colon\mathcal{H}\to\mathcal{H}$ is a self-adjoint operator on $\mathcal{H}$ with eigenvalues bounded in $[0,1/\lambda]$. To simplify the notation, we define function class $\mathcal{F}(\lambda)$ as

$$\mathcal{F}(\lambda) = \Big\{\|\phi(\cdot)\|_\Upsilon = \big[\phi(\cdot)^\top\Upsilon\phi(\cdot)\big]^{1/2}\colon \|\Upsilon\|_{\mathrm{op}}\le 1/\lambda\Big\}, \tag{D.11}$$

where $\Upsilon\colon\mathcal{H}\to\mathcal{H}$ in (D.11) is a self-adjoint operator on $\mathcal{H}$ whose eigenvalues are all bounded by $1/\lambda$ in magnitude. Here, the operator norm of Υ is defined as

$$\|\Upsilon\|_{\mathrm{op}} = \sup\big\{f^\top\Upsilon f\colon f\in\mathcal{H}, \|f\|_{\mathcal{H}}=1\big\} = \sup\big\{\langle f,\Upsilon f\rangle_{\mathcal{H}}\colon f\in\mathcal{H},\|f\|_{\mathcal{H}}=1\big\}.$$

Thus, by definition, for any finite subset $\mathcal{D}$ of $\mathcal{Z}$, $\|\phi(\cdot)\|_{\Lambda_{\mathcal{D}}^{-1}}$ belongs to $\mathcal{F}(\lambda)$, where $\Lambda_{\mathcal{D}}$ is defined in (D.5). For any $\epsilon\in(0,1)$, we let $N_\infty(\epsilon,\mathcal{F},\lambda)$ denote the ϵ-covering number of $\mathcal{F}(\lambda)$ in (D.11) with respect to the ℓ_∞-norm. Besides, let $N_\infty(\epsilon, \mathcal{H}, R)$ denote the ϵ-covering number of $\{f \in \mathcal{H}\colon \|f\|_{\mathcal{H}} \le R\}$ with respect to the ℓ_∞-norm and let $N(\epsilon, B)$ denote the ϵ-covering number of the interval $[0, B]$ with respect the Euclidean distance. Then, by (D.10) we obtain that

$$\big|\mathcal{C}\big(\mathcal{Q}_{\mathrm{ucb}}(h, R, B), \epsilon\big)\big| \le N_\infty(\epsilon/3, \mathcal{H}, R) \cdot N(\epsilon \cdot \lambda/3, B) \cdot N_\infty\big(\epsilon/(3B), \mathcal{F}, \lambda\big). \tag{D.12}$$

As shown in Vershynin (2018, Corollary 4.2.13), it holds that

$$N(\epsilon \cdot \lambda/3, B) \le 1 + 6B/(\epsilon \cdot \lambda) \le 1 + 6B/\epsilon, \tag{D.13}$$

where the last inequality follows from the fact that $\lambda \in [1, 2]$.

It remains to bound the first and the third terms on the right-hand side of (D.12) separately. We establish the ℓ_∞-covering of the RKHS norm ball and $F(\lambda)$ in the following two lemmas, respectively.

Lemma D.2 (ℓ_∞-norm covering number of RKHS ball)**.** For any $\epsilon \in (0, 1)$, we let $N_\infty(\epsilon, \mathcal{H}, R)$ denote the ϵ-covering number of the RKHS norm ball $\{f \in \mathcal{H}\colon \|f\|_{\mathcal{H}} \le R\}$ with respect to the ℓ_∞-norm. Consider the three eigenvalue decay conditions given in Assumption 4.3. Then, under Assumption 4.3, there exist absolute constants C_3 and C_4 such that

$$\log N_\infty(\epsilon, \mathcal{H}, R) \le \begin{cases} C_3 \cdot \gamma \cdot \big[\log(R/\epsilon) + C_4\big] & \gamma\text{-finite spectrum,} \\ C_3 \cdot \big[\log(R/\epsilon) + C_4\big]^{1+1/\gamma} & \gamma\text{-exponential decay,} \\ C_3 \cdot (R/\epsilon)^{2/[\gamma \cdot (1-2\tau)-1]} \cdot \big[\log(R/\epsilon) + C_4\big] & \gamma\text{-polynomial decay,} \end{cases}$$

where C_3 and C_4 are independent of T, H, R, and ϵ, and only depend on absolute constants C_ψ, C_1, C_2, γ, and τ specified in Assumption 4.3.

Proof. See §E.2 for a detailed proof. $\square$

Lemma D.3. For any $\epsilon \in (0, 1/e)$, let $N_\infty(\epsilon, \mathcal{F}, \lambda)$ be the ϵ-covering number of function class $\mathcal{F}(\lambda)$ with respect to the ℓ_∞-norm, where $\mathcal{F}(\lambda)$ is defined in (D.11). Here we assume that λ is bounded in $[c_1, c_2]$, where both c_1 and c_2 are absolute constants. Then, under Assumption 4.3, there exist absolute constants C_5 and C_6 such that

$$\log N_\infty(\epsilon, \mathcal{F}, \lambda) \le \begin{cases} C_5 \cdot \gamma^2 \cdot \big[\log(1/\epsilon) + C_6\big] & \gamma\text{-finite spectrum,} \\ C_5 \cdot \big[\log(1/\epsilon) + C_6\big]^{1+2/\gamma} & \gamma\text{-exponential decay,} \\ C_5 \cdot (1/\epsilon)^{4/[\gamma \cdot (1-2\tau)-1]} \cdot \big[\log(1/\epsilon) + C_6\big] & \gamma\text{-polynomial decay,} \end{cases}$$

where C_5 and C_6 only depend on C_ψ, C_1, C_2, γ, τ, c_1, and c_2, and do not rely on T, H, or ϵ.

Proof. See §E.3 for a detailed proof. $\square$

Finally, we conclude the proof by combining Lemmas D.2 and D.3. Specifically, by (D.12) and (D.13), we have

$$\begin{aligned} \log\big|\mathcal{C}\big(\mathcal{Q}_{\mathrm{ucb}}(h, R, B), \epsilon\big)\big| &\le \log N_\infty(\epsilon/3, \mathcal{H}, R) + \log N(\epsilon \cdot \lambda/3, B) + \log N_\infty\big(\epsilon/(3B), \mathcal{F}, \lambda\big) \\ &\le \log\big[1 + 6B/(\epsilon \cdot \lambda)\big] + \log N_\infty(\epsilon/3, \mathcal{H}, R) + \log N_\infty\big(\epsilon/(3B), \mathcal{F}, \lambda\big). \end{aligned} \tag{D.14}$$

We consider the three eigenvalue decay conditions separately. For the γ-finite spectrum case, by Lemmas D.2 and D.3 and (D.14) we have

$$
\begin{aligned}
&\log\big|\mathcal{C}\big(\mathcal{Q}_{\rm ucb}(h,R,B),\epsilon\big)\big| \\
&\qquad \le \log\big[1+6B/(\epsilon\cdot\lambda)\big] + C_3\cdot\gamma\cdot\big[\log(3R/\epsilon)+C_4\big] + C_5\cdot\gamma^2\cdot\big[\log(3B/\epsilon)+C_6\big] \\
&\qquad \le C_N\cdot\gamma\cdot\big[1+\log(R/\epsilon)\big] + C_N\cdot\gamma^2\cdot\big[1+\log(B/\epsilon)\big],
\end{aligned}
$$

where C_N is an absolute constant. Similarly, for the case where the eigenvalues satisfy the γ-exponential decay condition, by Lemmas D.2 and D.3 we have

$$
\begin{aligned}
&\log\big|\mathcal{C}\big(\mathcal{Q}_{\rm ucb}(h,R,B),\epsilon\big)\big| \\
&\qquad \le \log\big[1+6B/(\epsilon\cdot\lambda)\big] + C_3\cdot\big[\log(3R/\epsilon)+C_4\big]^{1+1/\gamma} + C_5\cdot\big[\log(3B/\epsilon)+C_6\big]^{1+2/\gamma} \\
&\qquad \le C_N\cdot\big[1+\log(R/\epsilon)\big]^{1+1/\gamma} + C_N\cdot\big[1+\log(B/\epsilon)\big]^{1+2/\gamma}
\end{aligned}
$$

for some absolute constant $C_N>0$. Finally, for the case of γ-polynomial eigenvalue decay, we have

$$
\begin{aligned}
&\log\big|\mathcal{C}\big(\mathcal{Q}_{\rm ucb}(h,R,B),\epsilon\big)\big| \\
&\qquad \le \log\big[1+6B/(\epsilon\cdot\lambda)\big] + C_3\cdot(3R/\epsilon)^{2/[\gamma\cdot(1-2\tau)-1]}\cdot\big[\log(3R/\epsilon)+C_4\big] \\
&\qquad\qquad + C_5\cdot(3B/\epsilon)^{4/[\gamma\cdot(1-2\tau)-1]}\cdot\big[\log(3B/\epsilon)+C_6\big] \\
&\qquad \le C_N\cdot(R/\epsilon)^{2/[\gamma\cdot(1-2\tau)-1]}\cdot\big[1+\log(R/\epsilon)\big] + C_N\cdot(B/\epsilon)^{4/[\gamma\cdot(1-2\tau)-1]}\cdot\big[1+\log(B/\epsilon)\big],
\end{aligned}
$$

where $C_N>0$ is an absolute constant. Therefore, we conclude the proof. □

D.2 Effective Dimension of RKHS

Definition D.4 (Maximal information gain)**.** For any fixed integer T and any $\sigma>0$, we define the maximal information gain associated with the RKHS $\mathcal{H}$ as

$$
\Gamma_K(T,\sigma^2) = \sup_{\mathcal{D}\subseteq\mathcal{Z}}\big\{1/2\cdot\mathrm{logdet}(I+\sigma^{-2}\cdot K_{\mathcal{D}})\big\}, \tag{D.15}
$$

where the supremum is taken over all discrete subsets of $\mathcal{Z}$ with cardinality no more than T, and K_D is the Gram matrix induced by $\mathcal{D}\subseteq\mathcal{Z}$, which is defined similarly as in (3.7). Here the subscript K in $\Gamma_K(T,\sigma^2)$ denotes the kernel function of $\mathcal{H}$.

The maximal information gain naturally arises in Gaussian process regression. Specifically, let $f\sim \mathrm{GP}(0,K)$ be a sample from the Gaussian process with covariance kernel K. Let $\mathcal{D}=\{z_1,\ldots,z_{|\mathcal{D}|}\}$ be a subset of $\mathcal{Z}$ with $|\mathcal{D}|\le T$ elements. Suppose that we observe noisy observations of f at points in $\mathcal{D}$. That is, for any $z_i\in\mathcal{D}$, we have $y_i=f(z_i)+\epsilon_i$, where $\epsilon_i\sim N(0,\sigma^2)$ is an i.i.d. random Gaussian noise. We let $y_{\mathcal{D}}$ denote the vector whose entries are y_i. Then, the information gain of $y_{\mathcal{D}}$ is defined as the mutual information between f and the observations $y_{\mathcal{D}}$, denoted by $I(f,y_{\mathcal{D}})$. By direct computation, we have

$$
I(f,y_{\mathcal{D}}) = 1/2\cdot\mathrm{logdet}(I+\sigma^{-2}\cdot K_{\mathcal{D}}).
$$

The mutual information $I(f, y_{\mathcal{D}})$ quantifies the reduction of the uncertainty about f when we observe $y_{\mathcal{D}}$. Thus, the maximal mutual information $\Gamma_K(T, \sigma^2)$ characterizes the maximal possible reduction of the uncertainty of f when having no more than T observations.

Moreover, we note that, when σ^2 is a constant, $\Gamma_K(T, \sigma^2)$ depends on the eigenvalue decay of the RKHS and thus can be viewed as an effective of the RKHS. Specifically, as shown in Srinivas et al. (2009), when the kernel is the d-dimensional linear kernel, $\Gamma_K(T, \sigma^2) = \mathcal{O}(d \log T)$. Moreover, for the squared exponential kernel and the Matérn kernel which have exponential and polynomial eigenvalue decay respectively, the maximal information gains are

$$\mathcal{O}\big((\log T)^{d+1}\big) \qquad \text{and} \qquad \mathcal{O}\big(T^{d(d+1)/(2\nu+d(d+1))} \cdot \log T\big),$$

respectively, where ν is the parameter of the Matérn kernel. In the following lemma, similar to Theorem 5 in Srinivas et al. (2009), we establish upper bounds on the maximal information gain of the RKHS under the three eigenvalue decay conditions specified in Assumption 4.3.

Lemma D.5 (Theorem 5 in Srinivas et al. (2009))**.** Let $\mathcal{Z}$ be a compact subset of $\mathbb{R}^d$ and $K\colon \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ be the RKHS kernel of $\mathcal{H}$. We assume that K is a bounded kernel in the sense that $\sup_{z \in \mathcal{Z}} K(z, z) \leq 1$, and K is continuously differentiable on $\mathcal{Z} \times \mathcal{Z}$. Moreover, let T_K be the integral operator induced by K and the Lebesgue measure on $\mathcal{Z}$, whose definition is given in (2.3). Let $\{\sigma_j\}_{j \geq 1}$ be the eigenvalues of T_K in descending order. We assume that $\{\sigma_j\}_{j \geq 1}$ satisfy either one of the following three eigenvalue decay conditions:

(i) γ-finite spectrum: We have $\sigma_j = 0$ for all $j \geq \gamma + 1$, where γ is a positive integer.

(ii) γ-exponential eigenvalue decay: There exist constants $C_1, C_2 > 0$ such that $\sigma_j \leq C_1 \exp(-C_2 \cdot j^{\gamma})$ for all $j \geq 1$, where $\gamma > 0$ is positive constant.

(iii) γ-polynomial eigenvalue decay: There exists a constant C_1 such that $\sigma_j \geq C_1 \cdot j^{-\gamma}$ for all $j \geq 1$, where $\gamma \geq 2 + 1/d$ is a constant.

Let σ be bounded in interval $[c_1, c_2]$ with c_1 and c_2 being absolute constants. Then, for conditions (i)–(iii) respectively, we have

$$\Gamma_K(T, \sigma^2) \leq \begin{cases} C_K \cdot \gamma \cdot \log T & \gamma\text{-finite spectrum}, \\ C_K \cdot (\log T)^{1+1/\gamma} & \gamma\text{-exponential decay}, \\ C_K \cdot T^{(d+1)/(\gamma+d)} \cdot \log T & \gamma\text{-polynomial decay}, \end{cases}$$

where C_K is an absolute constant that depends on d, γ, C_1, C_2, C, c_1, and c_2.

We note that Lemma D.5 is a generalization of Theorem 5 in Srinivas et al. (2009), which establishes the maximal information gain for the linear, squared exponential, and Matérn kernels, respectively. Specifically, the squared exponential kernel satisfies the γ-exponential eigenvalue decay condition with $\gamma = 1/d$. Lemma D.5 implies that the $\Gamma_K(T, \sigma^2) = \mathcal{O}((\log T)^{d+1})$, which matches Theorem 5 in Srinivas et al. (2009). Furthermore, the Matérn kernel with parameter ν satisfies the γ-polynomial eigenvalue decay condition with $\gamma = (2\nu + d)/d$. Then, Lemma D.5 asserts that $\Gamma_K(T, \sigma^2) = \mathcal{O}(T^{d(d+1)/(2\nu+d(d+1))} \cdot \log T)$, which also matches Theorem 5 in Srinivas et al. (2009).

Proof. The proof of this lemma is based on a modification of that of Theorem 5 in Srinivas et al. (2009). To begin with, for any $j \in \mathbb{N}$, we define $B_K(j) = \sum_{s>j} \sigma_s$, i.e., the sum of eigenvalues with indices larger than j. Then, we use the following lemma obtained from Srinivas et al. (2009) to bound $\Gamma_K(T, \sigma^2)$ using function B_K.

Lemma D.6 (Theorem 8 in Srinivas et al. (2009))**.** Under the same condition as in Lemma D.5, for any fixed $\tau > 0$, we denote $C_\tau = 2\mu(\mathcal{Z}) \cdot (2\tau + 1)$ where $\mu(\mathcal{Z})$ is the Lebesgue measure of $\mathcal{Z}$. Let n_T denote $C_\tau \cdot T^\tau \cdot \log T$. Then, for any $T_\star \in \{1, \ldots, n_T\}$, we have

$$\Gamma_K(T, \sigma^2) \le T_\star \cdot \log(T \cdot n_T/\sigma^2) + C_\tau \cdot \sigma^{-2} \cdot \log T \cdot \left[T^{\tau+1} \cdot B_K(T_\star) + 1\right] + \mathcal{O}(T^{1-\tau/d}).$$

Proof. See Srinivas et al. (2009) for a detailed proof. $\square$

In the following, we choose proper τ and $T_\star$ in Lemma D.6 for the three eigenvalue decay conditions separately.

Case (i): γ-Finite Spectrum. When $\sigma_j = 0$ for all $j \ge \gamma + 1$, we set $\tau = d$ and $T_\star = \gamma$ in Lemma D.6. Then we have $B_K(T_\star) = 0$ and $n_T = C_d \cdot T^d \cdot \log T$. When T is sufficiently large, it holds that $T_\star < n_T$. Then Lemma D.6 implies that

$$\Gamma_K(T, \sigma^2) \le \gamma \cdot \log\left(C_d \cdot T^{d+1} \cdot \log T/\sigma^2\right) + C_d \cdot \sigma^{-2} \cdot \log T + \mathcal{O}(1) \le C_K \cdot \gamma \cdot \log T$$

for some absolute constant $C_K > 0$. Thus, we conclude the proof for the first case.

Case (ii): γ-Exponential Decay. When $\{\sigma_j\}_{j\ge1}$ satisfies the γ-exponential eigenvalue decay condition, for any $T_\star \in \mathbb{N}$, we have

$$B_K(T_\star) = \sum_{j>T_\star} \sigma_j \le C_1 \cdot \sum_{j>T_\star} \exp(-C_2 \cdot j^\gamma) \le C_1 \cdot \int_{T_\star}^{\infty} \exp(-C_2 \cdot u^\gamma)\, \mathrm{d}u \tag{D.16}$$

Similar to the derivation of (E.16), by direct computation, we have

$$\int_{T_\star}^{\infty} \exp(-C_2 \cdot u^\gamma)\, \mathrm{d}u \le \begin{cases} C_2^{-1} \cdot \exp(-C_2 \cdot T_\star^\gamma), & \text{if } \gamma \ge 1, \\ 2 \cdot (\gamma \cdot C_2)^{-1} \cdot \exp(-C_2 \cdot T_\star^\gamma) \cdot T_\star^{1-\gamma}, & \text{if } \gamma \in (0, 1). \end{cases} \tag{D.17}$$

In the following, we set $\tau = d$. Then we have $n_T = C_d \cdot T^d \cdot \log T$ where $C_d = 2\mu(\mathcal{Z}) \cdot (2d + 1)$. Then we have

$$\log(T \cdot n_T) = \log(C_d) + \log \cdot \left(T^{d+1} \cdot \log T\right) \le \log(C_d) + 2(d+1) \cdot \log T \tag{D.18}$$

when T is sufficiently large. Moreover, combining Lemma D.6 and (D.18), when σ is sandwiched by absolute constants c_1 and c_2, we have

$$\Gamma_K(T, \sigma^2) \le \widetilde{C}_1 \cdot T_\star \cdot \log T + \widetilde{C}_2 \cdot \log T \cdot \left[T^{d+1} \cdot B_K(T_\star) + 1\right] + \widetilde{C}_3, \tag{D.19}$$

where $\widetilde{C}_1$, $\widetilde{C}_2$, and $\widetilde{C}_3$ are absolute constants that depend on d, γ, c_1, c_2, C_1, and C_2. Now we choose $T_\star$ such that

$$\exp(C_2 \cdot T_\star^\gamma) \asymp T \cdot n_T = C_d \cdot T^{d+1} \cdot \log T, \tag{D.20}$$

that is, $T_\star = \widetilde{C}_4 \cdot (\log T)^{1/\gamma}$ where $\widetilde{C}_4$ is an absolute constant. Notice that $T_\star < n_T$ when T is sufficiently large.

Thus, combining (D.16), (D.17), and (D.20), for $\gamma \geq 1$, we have

$$\begin{aligned} &\log T \cdot \big[T^{d+1} \cdot B_K(T_\star) + 1\big] \\ &\qquad \leq C_1 \cdot C_2^{-1} \log T \cdot T^{d+1} \cdot \exp(-C_2 \cdot T_\star^\gamma) + \log T \leq 2 \log T, \end{aligned} \tag{D.21}$$

where the last inequality follows from (D.20). Similarly, for $\gamma \in (0,1)$, by (D.16), (D.17), and (D.20), we have

$$\begin{aligned} &\log T \cdot \big[T^{d+1} \cdot B_K(T_\star) + 1\big] \\ &\qquad \leq 2C_1 \cdot (\gamma \cdot C_2)^{-1} \cdot \exp(-C_2 \cdot T_\star^\gamma) \cdot \log T \cdot T^{d+1} \cdot T_\star^{1-\gamma} + \log T \asymp (\log T)^{1/\gamma - 1} + \log T. \end{aligned} \tag{D.22}$$

Thus, combining (D.19), (D.21), (D.22), we conclude that

$$\Gamma_K(T, \sigma^2) \leq C_K \cdot \log(T)^{1+1/\gamma}$$

for any $\gamma \geq 0$, where C_K is an absolute constant that depends on d, γ, c_1, c_2, C_1, and C_2. Thus, we conclude the proof for the second case.

Case (iii): γ-Polynomial Decay. Similar to the former case, when $\{\sigma_j\}_{j\geq 1}$ satisfies the γ-polynomial eigenvalue decay condition, for any $T_\star \in \mathbb{N}$, we have

$$B_K(T_\star) = \sum_{j > T_\star} \sigma_j \leq C \cdot \sum_{j > T_\star} j^{-\gamma} \leq C \cdot \int_{T_\star}^\infty u^{-\gamma} \, \mathrm{d}u = \frac{C}{\gamma - 1} \cdot T_\star^{-(\gamma-1)}. \tag{D.23}$$

For a fixed $\tau \in (0, \gamma + d)$ to be determined later, recall that we denote $n_T = T^\tau \cdot \log T$. Hence, we have

$$\log(T \cdot n_T) = (1 + \tau) \cdot \log T + \log\log T \leq (\gamma + d + 2) \cdot \log T.$$

Combining Lemma D.6 and (D.23), we have

$$\Gamma_K(T, \sigma^2) \leq \widetilde{C}_5 \cdot T_\star \cdot \log T + \widetilde{C}_6 \cdot \log T \big[T^{\tau+1} \cdot T_\star^{-(\gamma-1)} + 1\big] + \widetilde{C}_7 \cdot T^{1-\tau/d}, \tag{D.24}$$

where $T_\star \in \{1, \ldots, n_T\}$, and $\widetilde{C}_5$, $\widetilde{C}_6$, $\widetilde{C}_7$ are absolute constants that depend on d, γ, c_1, c_2, and C.

Now we choose a $T_\star$ that balances the first two terms on the right-hand side of (D.24). Specifically, we let $T_\star = T^{(\tau+1)/\gamma}$. Then we have $T_\star = T^{\tau+1} \cdot T_\star^{-(\gamma-1)}$. Finally, solving the equation $T_\star = T^{1-\tau/d}$, we have $\tau = (\gamma - 1) \cdot d/(d + \gamma)$. Hence, by (D.24) we conclude that

$$\Gamma_K(T, \sigma^2) \leq C_K \cdot T_\star \cdot \log T = C_K \cdot T^{(d+1)/(\gamma+d)} \cdot \log T,$$

where C_K is an absolute constant. It remains to verify that $T_\star \leq n_T = T^\tau \cdot \log T$. This is true as long as $T_\star \leq T^\tau$, i.e., $\gamma \geq 2 + 1/d$. Therefore, we conclude the proof of Lemma D.5. $\square$

E Proofs of Auxiliary Results

In this section, we provide the proofs of the auxiliary results.

E.1 Proof of Lemma C.1

Proof. For any function $f \in \mathcal{H}$, using the feature representation induced by the kernel K, we have

$$
\begin{aligned}
\big|\langle f, \widehat{\theta}_h^t \rangle_{\mathcal{H}}\big| &= \big| f^\top \widehat{\theta}_h^t \big| \le \big| f^\top (\Lambda_h^t)^{-1} \Phi^\top y_h^t \big| \\
&= \Bigg| f^\top (\Lambda_h^t)^{-1} \sum_{\tau=1}^{t-1} \phi(x_h^\tau, a_h^\tau) \cdot [r_h(x_h^\tau, a_h^\tau) + V_{h+1}^t(x_{h+1}^\tau)] \Bigg|, \qquad \text{(E.1)}
\end{aligned}
$$

where we let Φ denote Φ_h^t defined in (C.13) for simplicity. Since $|r_h(x_h^\tau, a_h^\tau)| \le 1$ and $|V_{h+1}^t(x_{h+1}^\tau)| \le H - h$, we have $|[r_h(x_h^\tau, a_h^\tau) + V_{h+1}^t(x_{h+1}^\tau)]| \le H$ for all $h \in [H]$ and $\tau \in [t-1]$. Then, by (E.1) and Cauchy-Schwarz inequality, we have

$$
\begin{aligned}
\big|\langle f, \widehat{\theta}_h^t \rangle_{\mathcal{H}}\big| &\le H \cdot \sum_{\tau=1}^{t-1} \big| f^\top (\Lambda_h^t)^{-1} \phi(x_h^\tau, a_h^\tau) \big| \\
&\le H \cdot \Bigg[\sum_{\tau=1}^{t-1} f^\top (\Lambda_h^t)^{-1} f \Bigg]^{1/2} \cdot \Bigg[\sum_{\tau=1}^{t-1} \phi(x_h^\tau, a_h^\tau)^\top (\Lambda_h^t)^{-1} \phi(x_h^\tau, a_h^\tau) \Bigg]^{1/2} \\
&\le H/\sqrt{\lambda} \cdot \|f\|_{\mathcal{H}} \cdot \Bigg[\sum_{\tau=1}^{t-1} \phi(x_h^\tau, a_h^\tau)^\top (\Lambda_h^t)^{-1} \phi(x_h^\tau, a_h^\tau) \Bigg]^{1/2}, \qquad \text{(E.2)}
\end{aligned}
$$

where the last inequality follows from the fact that $(\Lambda_h^t)^{-1} \colon \mathcal{H} \to \mathcal{H}$ is a self-adjoint and positive definite operator whose eigenvalues are bounded by $1/\lambda$. Furthermore, by Lemma E.3, we have

$$
\Bigg[\sum_{\tau=1}^{t-1} \phi(x_h^\tau, a_h^\tau)^\top (\Lambda_h^t)^{-1} \phi(x_h^\tau, a_h^\tau) \Bigg] \le 2\mathrm{logdet}(I + K_h^t/\lambda), \qquad \text{(E.3)}
$$

Thus, combining (E.2), (E.3), and the fact that $\lambda \ge 1$, we obtain that

$$
\big|\langle f, \widehat{\theta}_h^t \rangle_{\mathcal{H}}\big| \le H \cdot \|f\|_{\mathcal{H}} \cdot \sqrt{2/\lambda \cdot \mathrm{logdet}(I + K_h^t/\lambda)} \le H \cdot \|f\|_{\mathcal{H}} \cdot \sqrt{2 \cdot \mathrm{logdet}(I + K_h^t/\lambda)}.
$$

Finally, utilizing the definition of $\Gamma_K(T, \lambda)$ in (D.15), we conclude the proof of this lemma. ☐

E.2 Proof of Lemma D.2

Proof. Recall that we have defined the integral operator $T_K \colon \mathcal{L}^2(\mathcal{Z}) \to \mathcal{L}^2(\mathcal{Z})$ defined in (2.3), which has eigenvalues $\{\sigma_j\}_{j \ge 0}$ and eigenvectors $\{\psi_j\}_{j \ge 0}$. Moreover, $\{\psi_j\}$ and $\{\sqrt{\sigma_j} \cdot \psi_j\}_{j \ge 0}$ are orthonormal bases of $\mathcal{L}_2(\mathcal{Z})$ and $\mathcal{H}$, respectively. Then, any $\in \mathcal{H}$ with $\|f\|_{\mathcal{H}} \le R$ can be written as

$$
f = \sum_{j=1}^{\infty} w_j \cdot \sqrt{\sigma_j} \cdot \psi_j, \qquad \text{(E.4)}
$$

where $\{w_j\}_{j \ge 0}$ satisfy $\sum_{j=1}^{\infty} w_j^2 = \|f\|_{\mathcal{H}}^2 \le R^2$. Let m be any positive integer and let $\Pi_m \colon \mathcal{H} \to \mathcal{H}$ denote the projection onto the subspace spanned by $\{\psi_j\}_{j \in [m]}$, i.e., $\Pi_m(f) = \sum_{j=1}^{m} w_j \cdot \sqrt{\sigma_j} \cdot \psi_j$ for any $f \in \mathcal{H}$ written as in (E.4). Then we have

$$
\|f - \Pi_m(f)\|_\infty = \sum_{j=m+1}^{\infty} |w_j| \cdot \sqrt{\sigma_j} \cdot \sup_{z \in \mathcal{Z}} |\psi_j(z)|. \qquad \text{(E.5)}
$$

In the following, we consider the three eigenvalue decay conditions specified in Assumption 4.3 separately.

Case (i): γ-Finite Spectrum. Consider the case where $\sigma_j = 0$ for all $j > \gamma$. Then, by the definition of Π_m, we have $f = \Pi_\gamma(f)$ for all $f \in \mathcal{H}$. That is, (E.4) is reduced to

$$f = \sum_{j=1}^{\gamma} w_j \cdot \sqrt{\sigma_j} \cdot \psi_j.$$

where $\{w_j\}_{j\in[\gamma]}$ satisfies $\sum_{j=1}^{\gamma} w_j^2 \le R^2$. Let $\mathcal{C}_\gamma(\epsilon, R)$ be the minimal ϵ-cover of the γ-dimensional Euclidean ball $\{w \in \mathbb{R}^\gamma : \|w\|_2 \le R\}$ with respect to the Euclidean norm. Then, by construction, there exists $\widetilde{w} \in \mathbb{R}^\gamma$ such that $\sum_{j=1}^{\gamma}(w_j - \widetilde{w}_j)^2 \le \epsilon^2$. Then, by Cauchy-Schwarz inequality, we have

$$\begin{aligned}\Bigl\| f - \sum_{j=1}^{\gamma} \widetilde{w}_j \cdot \sqrt{\sigma_j} \cdot \psi_j \Bigr\|_\infty &= \sup_{z\in\mathcal{Z}} \Bigl| \sum_{j=1}^{\gamma} (w_j - \widetilde{w}_j) \cdot \sqrt{\sigma_j} \cdot \psi_j(z) \Bigr| \\ &= \Bigl[\sum_{j=1}^{\gamma} (w_j - \widetilde{w}_j)^2 \Bigr]^{1/2} \cdot \sup_{z\in\mathcal{Z}} \biggl\{ \Bigl[\sum_{j=1}^{\gamma} \cdot |\psi_j(z)|^2 \Bigr]^{1/2} \biggr\} \le \epsilon \cdot \sup_z \sqrt{K(z,z)} \le \epsilon,\end{aligned} \tag{E.6}$$

where the last equality follows from the fact that $K(z,z) = \sum_{j=1}^{\gamma} \sigma_j \cdot |\psi_j(z)|^2$. Thus, the ϵ-covering of $\{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \le R\}$ is bounded by the cardinality of $\mathcal{C}_\gamma(\epsilon, R)$. As shown in Vershynin (2018, Corollary 4.2.13), we have

$$\bigl|\mathcal{C}_\gamma(\epsilon, R)\bigr| \le (1 + 2R/\epsilon)^\gamma. \tag{E.7}$$

Thus, combining (E.6) and (E.7), we have

$$\log N_\infty(\epsilon, \mathcal{H}, R) \le \gamma \cdot \log(1 + 2R/\epsilon) \le C_3 \cdot \gamma \cdot \bigl[\log(R/\epsilon) + C_4\bigr],$$

where both C_3 and C_4 are absolute constants. Thus, we conclude the proof for the first case.

Case (ii): γ-Exponential Decay. In the following, we assume the eigenvalues $\{\sigma_j\}_{j\ge1}$ satisfy the γ-exponential decay condition and $\|\psi_j\|_\infty \le C_\psi \cdot \sigma_j^{-\tau}$ for all $j \ge 1$. Thus, by (E.5) we have

$$\begin{aligned}\|f - \Pi_m(f)\|_\infty &\le \sum_{j=m+1}^{\infty} C_\psi \cdot |w_j| \cdot \sigma_j^{1/2-\tau} \\ &\le \sum_{j=m+1}^{\infty} C_\psi \cdot C_1^{1/2-\tau} \cdot |w_j| \cdot \exp\bigl[-C_2 \cdot (1/2 - \tau) \cdot j^\gamma\bigr].\end{aligned} \tag{E.8}$$

To simplify the notation, we define $C_{1,\tau} = C_\psi \cdot C_1^{1/2-\tau}$ and $C_{2,\tau} = C_2 \cdot (1 - 2\tau)$. Then, applying Cauchy-Schwarz inequality to (E.8), we have

$$\begin{aligned}\|f - \Pi_m(f)\|_\infty &\le C_{1,\tau} \cdot \Bigl(\sum_{j=m+1}^{\infty} |w_j|^2 \Bigr)^{1/2} \cdot \Bigl[\sum_{j=m+1}^{\infty} \exp(-C_{2,\tau} \cdot j^\gamma) \Bigr]^{1/2} \\ &\le C_{1,\tau} \cdot R \cdot \Bigl[\sum_{j=m+1}^{\infty} \exp(-C_{2,\tau} \cdot j^\gamma) \Bigr]^{1/2},\end{aligned} \tag{E.9}$$

where the second inequality follows from the fact that $\sum_{j\geq 1} w_j^2 \leq R^2$. Since $\gamma > 0$, $\exp(-u^\gamma)$ is monotonically decreasing in u. Thus, we have

$$\sum_{j=m+1}^{\infty} \exp(-C_{2,\tau} \cdot j^\gamma) \leq \int_m^\infty \exp(-C_{2,\tau} \cdot u^\gamma)\, \mathrm{d}u. \tag{E.10}$$

In the following, we bound the integral in (E.10) by considering the cases where $\gamma \geq 1$ and $\gamma \in (0,1)$ separately. First, when $\gamma \geq 1$, since $d \geq 1$, we have $u^{\gamma-1} \geq 1$ for all $u \geq d$. Hence, we have

$$\begin{aligned}\int_m^\infty \exp(-C_{2,\tau} \cdot u^\gamma)\, \mathrm{d}u &\leq \int_m^\infty u^{\gamma-1} \cdot \exp(-C_{2,\tau} \cdot u^\gamma)\, \mathrm{d}u \\ &\leq \int_{m^\gamma}^\infty \exp(-C_{2,\tau} \cdot v)\, \mathrm{d}v = C_{2,\tau}^{-1} \cdot \exp(-C_{2,\tau} \cdot m^\gamma),\end{aligned} \tag{E.11}$$

where the second inequality follows from the change of variable $v = u^\gamma$ and the fact that $\gamma \geq 1$. Second, when $\gamma < 1$, by letting $v = u^\gamma$, we have

$$\begin{aligned}\int_m^\infty \exp(-C_{2,\tau} \cdot u^\gamma)\, \mathrm{d}u &= \frac{1}{\gamma} \cdot \int_{m^\gamma}^\infty \exp(-C_{2,\tau} \cdot v) \cdot v^{1/\gamma-1}\, \mathrm{d}v = \frac{1}{\gamma \cdot C_{2,\tau}} \int_{m^\gamma}^\infty v^{1/\gamma-1}\, \mathrm{d}[-\exp(-C_{2,\tau} \cdot v)] \\ &= \frac{1}{\gamma \cdot C_{2,\tau}} \cdot \exp(-C_{2,\tau} \cdot m^\gamma) \cdot m^{1-\gamma} + \frac{(1-\gamma)}{\gamma^2 \cdot C_{2,\tau}} \int_{m^\gamma}^\infty \exp(-C_{2,\tau} \cdot v) \cdot v^{1/\gamma-2}\, \mathrm{d}v,\end{aligned} \tag{E.12}$$

where the last equality follows from integration by parts. Moreover, by direct calculation, we have

$$\begin{aligned}\frac{1}{\gamma} \int_{m^\gamma}^\infty \exp(-C_{2,\tau} \cdot v) \cdot v^{1/\gamma-2}\, \mathrm{d}v &\leq \frac{1}{m^\gamma} \cdot \frac{1}{\gamma} \int_{m^\gamma}^\infty \exp(-C_{2,\tau} \cdot v) \cdot v^{1/\gamma-1}\, \mathrm{d}v \\ &= \frac{1}{m^\gamma} \int_m^\infty \exp(-C_{2,\tau} \cdot u^\gamma)\, \mathrm{d}u,\end{aligned} \tag{E.13}$$

where the first inequality follows from the fact that $v \geq m^\gamma$ in the integral and the second equality follows from letting $u = v^{1/\gamma}$. Then, combining (E.12) and (E.13), we have

$$\begin{aligned}&\int_m^\infty \exp(-C_{2,\tau} \cdot u^\gamma)\, \mathrm{d}u \\ &\quad \leq \frac{1}{\gamma \cdot C_{2,\tau}} \cdot \exp(-C_{2,\tau} \cdot m^\gamma) \cdot m^{1-\gamma} + \frac{1/\gamma - 1}{C_{2,\tau} \cdot m^\gamma} \cdot \int_m^\infty \exp(-C_{2,\tau} \cdot u^\gamma)\, \mathrm{d}u.\end{aligned} \tag{E.14}$$

Thus, when m is sufficiently large such that $m^\gamma \cdot C_{2,\tau} > 2/\gamma - 2$, by (E.14) we have

$$\begin{aligned}\int_m^\infty \exp(-C_{2,\tau} \cdot u^\gamma)\, \mathrm{d}u &\leq \left(1 - \frac{1/\gamma - 1}{C_{2,\tau} m^\gamma}\right)^{-1} \cdot \frac{1}{\gamma \cdot C_{2,\tau}} \exp(-C_{2,\tau} \cdot m^\gamma) \cdot m^{1-\gamma} \\ &\leq \frac{2}{\gamma \cdot C_{2,\tau}} \exp(-C_{2,\tau} \cdot m^\gamma) \cdot m^{1-\gamma}.\end{aligned} \tag{E.15}$$

Therefore, combining (E.10), (E.11), and (E.15), we obtain that

$$\int_m^\infty \exp(-C_{2,\tau} \cdot u^\gamma)\, \mathrm{d}u \leq \begin{cases} C_{2,\tau}^{-1} \cdot \exp(-C_{2,\tau} \cdot m^\gamma), & \text{if } \gamma \geq 1, \\ 2 \cdot (\gamma \cdot C_{2,\tau})^{-1} \cdot \exp(-C_{2,\tau} \cdot m^\gamma) \cdot m^{1-\gamma}, & \text{if } \gamma \in (0,1). \end{cases} \tag{E.16}$$

In the sequel, we let m^* be the smallest integer such that

$$\int_m^\infty \exp(-C_{2,\tau} \cdot u^\gamma)\, \mathrm{d}u \le \left(\frac{\epsilon}{2C_{1,\tau} \cdot R}\right)^2, \qquad \forall m \ge m^*. \tag{E.17}$$

Hence, combining (E.9), (E.10), and (E.17), we have $\|f - \Pi_{m^*}(f)\|_\infty \le \epsilon/2$ for any $f \in \mathcal{H}$ with $\|f\|_{\mathcal{H}} \le R$. Besides, note that $C_{1,\tau}$, $C_{2,\tau}$, and γ are all absolute constants. By (E.16) and (E.17), there exist absolute constants $C_{1,m}$ and $C_{2,m}$ such that

$$m^* \le C_{1,m} \cdot \big[\log(R/\epsilon) + C_{2,m}\big]^{1/\gamma}. \tag{E.18}$$

Finally, it remains to approximate $\Pi_{m^*}(f)$ up to error $\epsilon/2$ for m^* specified in (E.17). By the expansion of f in (E.4), we have $\Pi_{m^*}(f) = \sum_{j=1}^{m^*} w_j \cdot \sqrt{\sigma_j} \cdot \psi_j$. For any m^* real numbers $\{\widetilde{w}_j\}_{j\in[m^*]}$, by Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\bigg|\big[\Pi_{m^*}(f)\big](z) - \sum_{j=1}^{m^*} \widetilde{w}_j \cdot \sqrt{\sigma_j} \cdot \psi_j(z)\bigg| &= \bigg|\sum_{j=1}^{m^*}(w_j - \widetilde{w}_j) \cdot \sqrt{\sigma_j} \cdot \psi_j(z)\bigg| \\
&\le \bigg[\sum_{j=1}^{m^*}(w_j - \widetilde{w}_j)^2\bigg]^{1/2} \cdot \bigg\{\sum_{j=1}^{m^*} \sigma_j \cdot [\psi_j(z)]^2\bigg\}^{1/2} \le \sqrt{K(z,z)} \cdot \bigg[\sum_{j=1}^{m^*}(w_j - \widetilde{w}_j)^2\bigg]^{1/2},
\end{aligned} \tag{E.19}$$

where the last inequality follows from the fact that $K(z,z) = \sum_{j=1}^\infty \sigma_j \cdot [\psi_j(z)]^2$. Under Assumption 4.3, we have $\sup_{z\in\mathcal{Z}} K(z,z) \le 1$. Notice that $\sum_{j=1}^{m^*} \omega_j^2 \le \|f\|_{\mathcal{H}}^2 \le R^2$. Let $\mathcal{C}_{m^*}(\epsilon/2, R)$ be the minimal $\epsilon/2$-cover of $\{w \in \mathbb{R}^{m^*} \colon \|w\|_2 \le R\}$ with respect to the Euclidean norm. By definition, for any $f \in \mathcal{H}$ with $\|f\|_{\mathcal{H}} \le R$, there exist $\widetilde{w} \in \mathcal{C}_{m^*}(\epsilon/2, R)$ such that $\sum_{j=1}^{m^*}(w_j - \widetilde{w}_j)^2 \le \epsilon^2/4$. Therefore, by (E.19) we have

$$\bigg\|f - \sum_{j=1}^{m^*} \widetilde{w}_j \cdot \sqrt{\sigma_j} \cdot \psi_j\bigg\|_\infty \le \|f - \Pi_{m^*}(f)\|_\infty + \bigg\|\Pi_{m^*}(f) - \sum_{j=1}^{m^*} \widetilde{w}_j \cdot \sqrt{\sigma_j} \cdot \psi_j\bigg\|_\infty \le \epsilon, \tag{E.20}$$

which implies that the ϵ-covering number of the RKHS norm ball $\{f \in \mathcal{H} \colon \|f\|_{\mathcal{H}} \le R\}$ is bounded by the cardinality of $\mathcal{C}_{m^*}(\epsilon/2, R)$, i.e., $N_\infty(\epsilon, \mathcal{H}, R) \le \big|\mathcal{C}_{m^*}(\epsilon/2, R)\big|$. As shown in Vershynin (2018, Corollary 4.2.13), we have

$$\big|\mathcal{C}_{m^*}(\epsilon/2, R)\big| \le (1 + 4R/\epsilon)^{m^*}. \tag{E.21}$$

Therefore, combining (E.18) and (E.21), we have

$$\begin{aligned}
\log N_\infty(\epsilon, \mathcal{H}, R) &\le m^* \cdot \log(1 + 4R/\epsilon) \le C_{1,m} \cdot \big[\log(R/\epsilon) + C_{2,m}\big]^{1/\gamma} \cdot [\log(1 + 4R/\epsilon)] \\
&\le C_3 \cdot \big[\log(R/\epsilon) + C_4\big]^{1+1/\gamma},
\end{aligned}$$

where C_3 and C_4 are absolute constants that only depend on C_Ψ, C_1, C_2, γ, and τ, which are specified in Assumption 4.3. Now we conclude the proof for the second case.

Case (iii): γ-Polynomial Decay. Finally, we consider the last case where the eigenvalues $\{\sigma_j\}_{j\geq 1}$ satisfy the γ-polynomial decay condition. The proof is similar to that of **Case (ii)**. Specifically, under Assumption 4.3 and by (E.5) we have

$$
\begin{aligned}
\|f - \Pi_m(f)\|_\infty &\leq \sum_{j=m+1}^{\infty} |w_j| \cdot \|\psi_j\|_\infty \leq \sum_{j=m+1}^{\infty} C_\psi \cdot |w_j| \cdot \sigma_j^{1/2-\tau} \\
&\leq \sum_{j=m+1}^{\infty} C_\psi \cdot C_1^{1/2-\tau} \cdot |w_j| \cdot j^{-\gamma\cdot(1/2-\tau)}
\end{aligned} \tag{E.22}
$$

for any $m \geq 1$. To simplify the notation, we define $C_{1,\tau} = C_\psi \cdot C_1^{1/2-\tau}$. Applying Cauchy-Schwarz inequality to (E.22) and using the fact that $\sum_{j\geq 1} |w_j|^2 \leq R^2$, we have

$$
\begin{aligned}
\|f - \Pi_m(f)\|_\infty &\leq C_{1,\tau} \cdot \biggl(\sum_{j=m+1}^{\infty} |w_j|^2\biggr)^{1/2} \cdot \biggl(\sum_{j=m+1}^{\infty} j^{-\gamma\cdot(1-2\tau)}\biggr)^{1/2} \\
&\leq C_{1,\tau} \cdot R \cdot \biggl(\int_m^\infty u^{-\gamma\cdot(1-2\tau)}\, \mathrm{d}u\biggr)^{1/2} = \frac{C_{1,\tau} \cdot R}{\sqrt{\gamma \cdot (1-2\tau) - 1}} \cdot m^{-[\gamma\cdot(1-2\tau)-1]/2}.
\end{aligned} \tag{E.23}
$$

We define m^* as the smallest integer such that the right-hand side of (E.23) is bounded by $\epsilon/2$. Notice that $C_{1,\tau}$, γ, and τ are absolute constants. Then, there exists an absolute constant $C_m > 0$ such that

$$
m^* \leq C_m \cdot (R/\epsilon)^{2/[\gamma\cdot(1-2\tau)-1]}. \tag{E.24}
$$

Furthermore, it remains to approximate $\Pi_{m^*}(f)$ up to error $\epsilon/2$. Similar to the previous case, we let $\mathcal{C}_{m^*}(\epsilon/2, R)$ be the minimal $\epsilon/2$-cover of the $\{w \in \mathbb{R}^{m^*} : \|w\|_2 \leq R\}$ with respect to the Euclidean norm. Then by definition, for any $f \in \mathcal{H}$ with $\|f\|_{\mathcal{H}} \leq R$, there exists $\widetilde{w} \in \mathcal{C}_{m^*}(\epsilon/2, R)$ such that (E.20) holds, which implies that

$$
\begin{aligned}
\log N_\infty(\epsilon, \mathcal{H}, R) &\leq \log \bigl|\mathcal{C}_{m^*}(\epsilon/2, R)\bigr| \leq m^* \cdot \log(1 + 4R/\epsilon) \\
&\leq C_3 \cdot (R/\epsilon)^{2/[\gamma\cdot(1-2\tau)-1]} \cdot \bigl[\log(R/\epsilon) + C_4\bigr],
\end{aligned}
$$

where C_3 and C_4 are absolute constants. Here the second inequality follows from Corollary 4.2.13 in Vershynin (2018) and the third inequality follows from (E.24). Therefore, we conclude the proof of this lemma. □

E.3 Proof of Lemma D.3

Proof. As shown in §2.2, the feature mapping $\phi: \mathcal{Z} \to \mathcal{H}$ satisfies

$$
\phi(z) = \sum_{j=1}^{\infty} \sigma_j \cdot \psi_j(z) \cdot \psi_j = \sum_{j=1}^{\infty} \sqrt{\sigma_j} \cdot \psi_j(z) \cdot (\sqrt{\sigma_j} \cdot \psi_j). \tag{E.25}
$$

That is, when expanding $\phi(z) \in \mathcal{H}$ in the basis $\{\sqrt{\sigma_j} \cdot \psi_j\}_{j \geq 0}$ as in (E.4), the j-th coefficient is equal to $\sqrt{\sigma_j} \cdot \psi_j(z)$ for all $j \geq 1$. Similar to the proof of Lemma D.2, in the following, we consider the three eigenvalue decay conditions separately.

Case (i): γ-Finite Spectrum. When $\mathcal{H}$ has only γ nonzero eigenvalues, for any $z \in \mathcal{Z}$, we define a vector $w_z \in \mathbb{R}^\gamma$ by letting its j-th entry be $\sqrt{\sigma_j} \cdot \psi_j(z)$ for all $j \in [\gamma]$. Moreover, for any self-adjoint operator $\Upsilon \colon \mathcal{H} \to \mathcal{H}$ satisfying $\|\Upsilon\|_{\mathrm{op}} \leq 1/\lambda$, we define a matrix $A_\Upsilon \in \mathbb{R}^{\gamma \times \gamma}$ as follows. For any $j, k \in [\gamma]$, we define the (j,k)-th entry of A_Υ as

$$[A_\Upsilon]_{j,k} = \bigl\langle \sqrt{\sigma_j} \cdot \psi_j, \sqrt{\sigma_k} \cdot \Upsilon \psi_k \bigr\rangle_{\mathcal{H}}.$$

By (E.25) and the definition of A_Υ, we have

$$\|\phi(z)\|_\Upsilon^2 = \sum_{j,k=1}^{\gamma} \sqrt{\sigma_j} \cdot \psi_j(z) \cdot \sqrt{\sigma_k} \cdot \psi_k(z) \cdot [A_\Upsilon]_{j,k} = w_z^\top A_\Upsilon w_z. \tag{E.26}$$

With a slight abuse of notation, we define $\mathcal{C}_\gamma(\epsilon, \lambda)$ denote the minimal ϵ^2-cover of

$$\bigl\{ A \in \mathbb{R}^{\gamma \times \gamma} \colon \|A\|_{\mathrm{fro}} \leq \sqrt{\gamma}/\lambda \bigr\}$$

with respect to the Frobenius norm. Then by definition, there exists $\widetilde{A}_\Upsilon \in \mathcal{C}_\gamma(\epsilon, \lambda)$ such that $\|A_\Upsilon - \widetilde{A}_\Upsilon\|_{\mathrm{fro}} \leq \epsilon^2$, which implies that

$$\bigl| w_z^\top A_\Upsilon w_z - w_z^\top \widetilde{A}_\Upsilon w_z \bigr| \leq \|w_z\|_2^2 \cdot \|A_\Upsilon - \widetilde{A}_\Upsilon\|_{\mathrm{op}} \leq \|A_\Upsilon - \widetilde{A}_\Upsilon\|_{\mathrm{fro}} \leq \epsilon^2, \tag{E.27}$$

where we use the fact that

$$\|w_z\|_2^2 = \sum_{j=1}^{\gamma} |w_j|^2 = \sum_{j=1}^{\gamma} \sigma_j \cdot |\psi_j(z)|^2 = K(z,z) \leq 1.$$

Thus, combining (E.26) and (E.27), and utilizing Corollary 4.2.13 in Vershynin (2018), we have

$$\log N_\infty(\epsilon, \mathcal{F}, \lambda) \leq \log \bigl| \mathcal{C}_\gamma(\epsilon, \lambda) \bigr| \leq \gamma^2 \cdot \log\bigl[1 + 8\sqrt{\gamma}/(\lambda \cdot \epsilon^2)\bigr] \leq C_5 \cdot \gamma^2 \cdot \bigl[\log(1/\epsilon) + C_6\bigr],$$

where C_5 and C_6 are absolute constants that depend solely on λ and γ. Thus, we conclude the proof for the first case.

Case (ii): γ-Exponential Decay. In the following, we focus on the second case where the eigenvalues satisfy the γ-exponential decay condition. For any $m \in \mathbb{N}$, we define $\Pi_m \colon \mathcal{H} \to \mathcal{H}$ as the projection operator onto the subspace spanned by $\{\psi_j\}_{j \in [m]}$. Then, by Cauchy-Schwarz inequality and Assumption 4.3, for any $z \in \mathcal{Z}$, by (E.25) we have

$$\begin{aligned} \bigl\| \phi(z) - \Pi_m\bigl[\phi(z)\bigr] \bigr\|_{\mathcal{H}} &= \Bigl\| \sum_{j=m+1}^{\infty} \sqrt{\sigma_j} \cdot \psi_j(z) \cdot \sqrt{\sigma_j} \cdot \psi_j \Bigr\|_{\mathcal{H}} = \Bigl\{ \sum_{j=m+1}^{\infty} \sigma_j \cdot [\psi_j(z)]^2 \Bigr\}^{1/2} \\ &\leq \Bigl(\sum_{j=m+1}^{\infty} \sigma_j \cdot \|\psi_j\|_\infty^2 \Bigr)^{1/2} \leq C_\psi \cdot \Bigl(\sum_{j=m+1}^{\infty} \sigma_j^{1-2\tau} \Bigr)^{1/2}, \end{aligned} \tag{E.28}$$

where the second equality follows from the fact that $\{\sqrt{\sigma_j}\cdot\psi_j\}_{j\geq 0}$ form an orthonormal basis of $\mathcal{H}$, the first inequality follows from taking supremum over $z\in\mathcal{Z}$, and the last inequality follows from the assumption that $\|\psi_j\|_\infty\leq C_\psi\cdot\sigma_j^{-\tau}$. Then, for any self-adjoint operator $\Upsilon\colon\mathcal{H}\to\mathcal{H}$ satisfying $\|\Upsilon\|_{\mathrm{op}}\leq 1/\lambda$ and any $z\in\mathcal{Z}$, by (E.28) and triangle inequality we have

$$\Big|\|\phi(z)\|_\Upsilon-\big\|\Pi_m\big[\phi(z)\big]\big\|_\Upsilon\Big|\leq\big\|\phi(z)-\Pi_m\big[\phi(z)\big]\big\|_\Upsilon\leq C_\psi/\sqrt{\lambda}\cdot\bigg(\sum_{j=m+1}^{\infty}\sigma_j^{1-2\tau}\bigg)^{1/2}. \tag{E.29}$$

Note that the eigenvalues $\{\sigma_j\}_{j\geq 0}$ admit γ-exponential decay under Assumption 4.3. Then we further upper bound the right-hand side of (E.29) by

$$\sup_{z\in\mathcal{Z}}\Big|\|\phi(z)\|_\Upsilon-\big\|\Pi_m\big[\phi(z)\big]\big\|_\Upsilon\Big|\leq C_\psi/\sqrt{\lambda}\cdot\bigg\{\sum_{j=m+1}^{\infty}C_1^{1-2\tau}\cdot\exp\big[-C_2\cdot(1-2\tau)\cdot j^\gamma\big]\bigg\}^{1/2}. \tag{E.30}$$

To simplify the notation, we define $C_{3,\tau}=C_\psi\cdot C_1^{1/2-\tau}/\sqrt{\lambda}$ and $C_{4,\tau}=C_2\cdot(1-2\tau)$, which are both absolute constants. Then, by (E.30) and the monotonicity of $\exp(-u^\gamma)$, we further obtain

$$\sup_{z\in\mathcal{Z}}\Big|\|\phi(z)\|_\Upsilon-\big\|\Pi_m\big[\phi(z)\big]\big\|_\Upsilon\Big|\leq C_{3,\tau}\cdot\bigg[\int_m^\infty\exp(-C_{4,\tau}\cdot u^\gamma)\,\mathrm{d}u\bigg]^{1/2}. \tag{E.31}$$

Here we can take the supremum over $\mathcal{Z}$ because the right-hand side of (E.30) does not depend on z. Note that we have shown in (E.16) that

$$\int_m^\infty\exp(-C_{4,\tau}\cdot u^\gamma)\,\mathrm{d}u\leq\begin{cases}C_{4,\tau}^{-1}\cdot\exp(-C_{4,\tau}\cdot m^\gamma), & \text{if }\gamma\geq 1,\\ 2\cdot(\gamma\cdot C_{4,\tau})^{-1}\cdot\exp(-C_{4,\tau}\cdot m^\gamma)\cdot m^{1/\gamma-1}, & \text{if }\gamma\in(0,1),\end{cases} \tag{E.32}$$

where for the case of $\gamma\in(0,1)$, (E.32) holds for sufficient large m such that $m^\gamma\cdot C_{4,\tau}>2/\gamma-2$.

Then, in the following, we define m^* as the smallest integer such that

$$\int_{m^*}^\infty\exp(-C_{4,\tau}\cdot u^\gamma)\,\mathrm{d}u\leq\big[\epsilon/(2C_{3,\tau})\big]^2. \tag{E.33}$$

By (E.32), since both $C_{3,\tau}$, $C_{4,\tau}$ and γ are absolute constants, there exist absolute constants $C_{3,m}$ and $C_{4,m}$ such that

$$m^*\leq C_{3,m}\cdot\big[\log(1/\epsilon)+C_{4,m}\big]^{1/\gamma}. \tag{E.34}$$

It is worth noting that the choice of m^* in (E.34) is uniform over all $z\in\mathcal{Z}$. Moreover, by (E.31), for such an m^*, it holds that

$$\sup_{z\in\mathcal{Z}}\Big|\|\phi(z)\|_\Upsilon-\big\|\Pi_{m^*}\big[\phi(z)\big]\big\|_\Upsilon\Big|\leq\epsilon/2. \tag{E.35}$$

Thus, it remains to approximate $\|\Pi_{m^*}[\phi(z)]\|_\Upsilon$ up to accuracy $\epsilon/2$. Note that the subspace spanned by $\{\psi_j\}_{j\in[m^*]}$ is m^* dimensional. When restricted to such a subspace, Υ can be expressed using a matrix $A_\Upsilon \in \mathbb{R}^{m^* \times m^*}$. Specifically, for any $j, k \in [m^*]$, we define the (j,k)-th entry of A_Υ as

$$[A_\Upsilon]_{j,k} = \big\langle \sqrt{\sigma_j} \cdot \psi_j, \sqrt{\sigma_k} \cdot \Upsilon \psi_k \big\rangle_{\mathcal{H}}. \tag{E.36}$$

Besides, let $w_z \in \mathbb{R}^{m^*}$ be a vector whose j-th entry is given by $\sqrt{\sigma_j} \cdot \psi_j(z)$, $\forall j \in [m^*]$. Then, by (E.36) it holds that

$$\big\|\Pi_{m^*}\big[\phi(z)\big]\big\|_\Upsilon^2 = \big\langle \Pi_{m^*}\big[\phi(z)\big], \Upsilon \Pi_{m^*}\big[\phi(z)\big] \big\rangle_{\mathcal{H}} = w_z^\top A_\Upsilon w_z. \tag{E.37}$$

Moreover, since $\|\Upsilon\|_{\mathrm{op}} \leq 1/\lambda$, the matrix operator norm of A_Υ is also bounded by $1/\lambda$, i.e., $\|A_\Upsilon\|_{\mathrm{op}} \leq 1/\lambda$. This means that the Frobenius norm of A_Υ is bounded by $\sqrt{m^*}/\lambda$. Let $\mathcal{C}_{m^*}(\epsilon/2, \lambda)$ denote the minimal $\epsilon^2/4$-cover of $\{A \in \mathbb{R}^{m^* \times m^*} \colon \|A\|_{\mathrm{fro}} \leq \sqrt{m^*}/\lambda\}$ with respect to the Frobenius norm. By definition, there exists $\widetilde{A}_\Upsilon \in \mathcal{C}_{m^*}(\epsilon/2, \lambda)$ such that $\|A_\Upsilon - \widetilde{A}_\Upsilon\|_{\mathrm{fro}} \leq \epsilon^2/4$. Hence, we have

$$\big|w_z^\top A_\Upsilon w_z - w_z^\top \widetilde{A}_\Upsilon w_z\big| \leq \|w_z\|_2^2 \cdot \|A_\Upsilon - \widetilde{A}_\Upsilon\|_{\mathrm{op}} \leq \|A_\Upsilon - \widetilde{A}_\Upsilon\|_{\mathrm{fro}} \leq \epsilon^2/4. \tag{E.38}$$

Finally, for any $z \in \mathcal{Z}$, we define

$$f_\Upsilon(z) = w_z^\top \widetilde{A}_\Upsilon w_z = \sum_{j,k=1}^{m^*} \sqrt{\sigma_j \cdot \sigma_k} \cdot \psi_j(z) \cdot \psi_k(z) \cdot \big[\widetilde{A}_\Upsilon\big]_{jk}, \tag{E.39}$$

where $[\widetilde{A}_\Upsilon]_{jk}$ is the (j,k)-th entry of $\widetilde{A}_\Upsilon$ and m^* is specified in (E.33). We remark that $f_\Upsilon \colon \mathcal{Z} \to \mathbb{R}$ is well-defined since m^* does not depend on z.

Finally, combining (E.35), (E.37), (E.38), and (E.39), we obtain

$$\begin{aligned}
\big\|\|\phi(z)\|_\Upsilon - f_\Upsilon\big\|_\infty &= \sup_{z \in \mathcal{Z}} \big|\|\phi(z)\|_\Upsilon - f_\Upsilon(z)\big| \\
&\leq \sup_{z \in \mathcal{Z}} \Big|\|\phi(z)\|_\Upsilon - \big\|\Pi_{m^*}\big[\phi(z)\big]\big\|_\Upsilon\Big| + \sup_{z \in \mathcal{Z}} \Big|\big\|\Pi_{m^*}\big[\phi(z)\big]\big\|_\Upsilon - f_\Upsilon(z)\Big| \\
&\leq \epsilon/2 + \sup_{z \in \mathcal{Z}} \Big|\sqrt{w_z^\top A_\Upsilon w_z} - \sqrt{w_z^\top \widetilde{A}_\Upsilon w_z}\Big| \leq \epsilon/2 + \sup_{z \in \mathcal{Z}} \sqrt{\big|w_z^\top A_\Upsilon w_z - w_z^\top \widetilde{A}_\Upsilon w_z\big|} \leq \epsilon.
\end{aligned}$$

This implies that $\{f_\Upsilon \colon \Upsilon \in \mathcal{C}_{m^*}(\epsilon, \lambda)\}$ forms an ϵ-cover of $\mathcal{F}(\lambda)$ in (D.11). Hence, it holds that

$$N_\infty(\epsilon, \mathcal{F}, \lambda) \leq \big|\mathcal{C}_{m^*}(\epsilon/2, \lambda)\big|. \tag{E.40}$$

Furthermore, using Corollary 4.2.13 in Vershynin (2018), we have

$$\big|\mathcal{C}_{m^*}(\epsilon/2, \lambda)\big| \leq \big[1 + 8\sqrt{m^*}/(\lambda \cdot \epsilon^2)\big]^{m^{*2}}. \tag{E.41}$$

Combining (E.34), (E.40), and (E.41), we finally have

$$\begin{aligned}
\log N_\infty(\epsilon, \mathcal{F}, \lambda) &\leq m^{*2} \cdot \log\big[1 + 8\sqrt{m^*}/(\lambda \cdot \epsilon^2)\big] \\
&\leq C_{3,m}^2 \cdot \big[\log(1/\epsilon) + C_{4,m}\big]^{2/\gamma} \cdot \log\big\{1 + 8C_{3,m}^{1/2} \cdot \big[\log(1/\epsilon) + C_{4,m}\big]^{1/(2\gamma)}/(\lambda \cdot \epsilon^2)\big\} \\
&\leq C_5 \cdot \big[\log(1/\epsilon) + C_6\big]^{1+2/\gamma},
\end{aligned}$$

where C_5 and C_6 are absolute constants that depend on C_ψ, C_1, C_2, τ, γ, and λ, but are independent of T, H, and ϵ. Here in the last inequality we use the fact that $\log(1/\epsilon) \le 1/\epsilon$, which holds when $\epsilon \le 1/e$. Therefore, we conclude the proof for the second case.

Case (iii): γ-Polynomial Decay. Finally, we consider the last case where the eigenvalues satisfy the γ-polynomial eigenvalue condition. Our proof is similar to that for the second case. Specifically, for any $m \ge 1$, by the assumption that $\sigma_j \le C_1 \cdot j^{-\gamma}$ for all $j \ge 1$, we have

$$\begin{aligned}
&\sup_{z\in\mathcal{Z}} \Big| \|\phi(z)\|_\Upsilon - \big\|\Pi_m[\phi(z)]\big\|_\Upsilon \Big| \\
&\quad \le C_\psi/\sqrt{\lambda} \cdot \bigg(\sum_{j=m+1}^{\infty} \sigma_j^{1-2\tau} \bigg)^{1/2} \le C_\psi \cdot C_1^{1/2-\tau}/\sqrt{\lambda} \cdot \bigg(\sum_{j=m+1}^{\infty} j^{-\gamma\cdot(1-2\tau)} \bigg)^{1/2} \\
&\quad \le C_\psi \cdot C_1^{1/2-\tau}/\sqrt{\lambda} \cdot \bigg(\int_m^\infty u^{-\gamma\cdot(1-2\tau)}\, \mathrm{d}u \bigg)^{1/2} = \frac{C_\psi \cdot C_1^{1/2-\tau}/\sqrt{\lambda}}{\sqrt{\gamma\cdot(1-2\tau)-1}} \cdot m^{-[\gamma\cdot(1-2\tau)-1]/2}. \qquad \text{(E.42)}
\end{aligned}$$

Notice that C_ψ, C_1, γ, and τ are all absolute constants. Similar to the previous case, we let m^* be the smallest integer m such that the

$$\sup_{z\in\mathcal{Z}} \Big| \|\phi(z)\|_\Upsilon - \big\|\Pi_m[\phi(z)]\big\|_\Upsilon \Big| \le \epsilon/2.$$

Then by (E.42), there exists an absolute constant $C_{3,m} > 0$ such that

$$m^* \le C_{3,m} \cdot (1/\epsilon)^{2/[\gamma\cdot(1-2\tau)-1]}. \qquad \text{(E.43)}$$

Recall that we let $C_{m^*}(\epsilon/2, \lambda)$ be the minimal $\epsilon^2/4$-cover of $\{A \in \mathbb{R}^{m^*\times m^*} : \|A\|_{\mathrm{fro}} \le \sqrt{m^*}/\lambda\}$ with respect to the Frobenius norm. As shown in (E.36) – (E.41), we similarly have

$$\begin{aligned}
\log N_\infty(\epsilon, \mathcal{F}, \lambda) &\le \log\big|\mathcal{C}_{m^*}(\epsilon/2,\lambda)\big| \le m^{*2} \cdot \log\big[1 + 8\sqrt{m^*}/(\lambda\cdot\epsilon^2)\big] \\
&\le C_{3,m}^2 \cdot (1/\epsilon)^{4/[\gamma\cdot(1-2\tau)-1]} \cdot \log\big[1 + 8\sqrt{m^*}/(\lambda\cdot\epsilon^2\big] \\
&\le C_5 \cdot (1/\epsilon)^{4/[\gamma\cdot(1-2\tau)-1]} \cdot \big[\log(1/\epsilon) + C_6\big],
\end{aligned}$$

where C_5 and C_6 are absolute constants that only depend on C_1, C_ψ, C_1, τ, γ, and λ. Here in the last inequality we utilize (E.43), which implies that

$$\sqrt{m^*}/(\lambda\cdot\epsilon^2) \le \sqrt{C_{3,m}}/\lambda \cdot (1/\epsilon)^{\frac{2\gamma\cdot(1-2\tau)-1}{\gamma\cdot(1-2\tau)-1}}.$$

Thus, we conclude the proof of the last case and therefore conclude the proof of the lemma. $\square$

E.4 Technical Lemmas

Next, we present a few concentration inequalities. The following one provides a concentration inequality for the standard self-normalized processes.

Lemma E.1 (Concentration of Self-Normalized Processes in RKHS (Chowdhury and Gopalan, 2017)). Let $\mathcal{H}$ be an RKHS defined over $\mathcal{X} \subseteq \mathbb{R}^d$ with kernel function $K(\cdot,\cdot)\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Let $\{x_\tau\}_{\tau=1}^\infty \subseteq \mathcal{X}$ be a discrete time stochastic process that is adapted to the filtration $\{\mathcal{F}_t\}_{t=0}^\infty$. That is, x_τ is $\mathcal{F}_{\tau-1}$ measurable for all $\tau \geq 1$. Let $\{\epsilon_t\}_{\tau=1}^\infty$ be a real-valued stochastic process such that (i) $\epsilon_\tau \in \mathcal{F}_\tau$ and (ii) ϵ_τ is zero-mean and σ-subGaussian conditioning on $\mathcal{F}_{\tau-1}$, i.e.,

$$\mathbb{E}[\epsilon_\tau|\mathcal{F}_{\tau-1}] = 0, \qquad \mathbb{E}[e^{\lambda\epsilon_\tau}|\mathcal{F}_{\tau-1}] \leq e^{\lambda^2\sigma^2/2}, \qquad \forall \lambda \in \mathbb{R}.$$

Moreover, for any $t \geq 2$, let $E_t = (\epsilon_1, \ldots, \epsilon_{t-1})^\top \in \mathbb{R}^{t-1}$ and $K_t \in \mathbb{R}^{(t-1)\times(t-1)}$ be the Gram matrix of $\{x_\tau\}_{\tau\in[t-1]}$. Then, for any $\eta > 0$ and any $\delta \in (0,1)$, with probability at least $1-\delta$, simultaneously for all $t \geq 1$, we have

$$E_t^\top \big[(K_t + \eta \cdot I)^{-1} + I\big]^{-1} E_t \leq \sigma^2 \cdot \mathrm{logdet}\big[(1+\eta)\cdot I + K_t\big] + 2\sigma^2 \cdot \log(1/\delta). \tag{E.44}$$

Moreover, if K_t is positive definite for all $t \geq 2$ with probability one, then the inequality in (E.44) also holds with $\eta = 0$.

Proof. See Theorem 1 in Chowdhury and Gopalan (2017) for a detailed proof. □

Lemma E.2 (Lemma D.4 of Jin et al. (2018)). Let $\{x_\tau\}_{\tau=1}^\infty$ and $\{\phi_\tau\}_{\tau=1}^\infty$ be an $\mathcal{S}$-valued and an $\mathcal{H}$-valued stochastic process adapted to filtration $\{\mathcal{F}_\tau\}_{\tau=0}^\infty$, respectively, where we assume that $\|\phi_\tau\|_\mathcal{H} \leq 1$ for all $\tau \geq 1$. Besides, for any $t \geq 1$, we let $K_t \in \mathbb{R}^{t\times t}$ be the Gram matrix of $\{\phi_\tau\}_{\tau\in[t]}$ and define an operator $\Lambda_t\colon \mathcal{H} \to \mathcal{H}$ as $\Lambda_t = \lambda\cdot I_\mathcal{H} + \sum_{\tau=1}^t \phi_\tau\phi_\tau^\top$ with $\lambda > 1$. Let $\mathcal{V} \subseteq \{V\colon \mathcal{S} \to [0,H]\}$ be a class of bounded functions on $\mathcal{S}$. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have simultaneously for all $t \geq 1$ that

$$\begin{aligned}&\sup_{V\in\mathcal{V}} \bigg\| \sum_{\tau=1}^t \phi_\tau \big\{V(x_\tau) - \mathbb{E}[V(x_\tau)|\mathcal{F}_{\tau-1}]\big\} \bigg\|_{\Lambda_t^{-1}}^2 \\ &\qquad \leq 2H^2 \cdot \mathrm{logdet}(I + K_t/\lambda) + 2H^2 t(\lambda-1) + 4H^2 \log(\mathcal{N}_\epsilon/\delta) + 8t^2\epsilon^2/\lambda\end{aligned} \tag{E.45}$$

where $\mathcal{N}_\epsilon$ is the ϵ-covering number of $\mathcal{V}$ with respect to the distance $\mathrm{dist}(\cdot,\cdot)$.

Proof. Let $\mathcal{V}_\epsilon \subseteq \{V\colon \mathcal{S} \to [0,H]\}$ be the minimal ϵ-cover of $\mathcal{V}$ such that $N_\epsilon = |\mathcal{V}_\epsilon|$. Then for any $V \in \mathcal{V}$, there exists a value function $V'\colon \mathcal{S} \to \mathbb{R}$ in $\mathcal{N}_\epsilon$ such that $\mathrm{dist}(V, V') \leq \epsilon$. Let $\Delta_V = V - V'$. By the inequality $(a+b)^2 \leq 2a^2 + 2b^2$, we have

$$\begin{aligned}&\bigg\| \sum_{\tau=1}^t \phi_\tau \big\{V(x_\tau) - \mathbb{E}[V(x_\tau)|\mathcal{F}_{\tau-1}]\big\} \bigg\|_{\Lambda_t^{-1}}^2 \\ &\quad \leq 2 \cdot \bigg\| \sum_{\tau=1}^t \phi_\tau \big\{V'(x_\tau) - \mathbb{E}[V'(x_\tau)|\mathcal{F}_{\tau-1}]\big\} \bigg\|_{\Lambda_t^{-1}}^2 + 2 \cdot \bigg\| \sum_{\tau=1}^t \phi_\tau \big\{\Delta_V(x_\tau) - \mathbb{E}[\Delta_V(x_\tau)|\mathcal{F}_{\tau-1}]\big\} \bigg\|_{\Lambda_t^{-1}}^2.\end{aligned} \tag{E.46}$$

To bound the first term on the right-hand side of (E.46), we apply Lemma E.1 to V' and take a union bound over $V' \in \mathcal{V}_\epsilon$. While for the second term, since $\sup_{x\in\mathcal{S}} |\Delta_V(x)| \leq \epsilon$, we have

$$\bigg\| \sum_{\tau=1}^t \phi_\tau \big\{\Delta_V(x_\tau) - \mathbb{E}[\Delta_V(x_\tau)|\mathcal{F}_{\tau-1}]\big\} \bigg\|_{\Lambda_t^{-1}}^2 \leq t^2 \cdot (2\epsilon)^2/\lambda = 4t^2\epsilon^2/\lambda. \tag{E.47}$$

Thus, combining (E.46) and (E.47), we have

$$
\begin{aligned}
&\sup_{V\in\mathcal{V}} \bigg\| \sum_{\tau=1}^{t} \phi_\tau \big\{ V(x_\tau) - \mathbb{E}[V(x_\tau)|\mathcal{F}_{\tau-1}] \big\} \bigg\|^2_{\Lambda_t^{-1}} \\
&\qquad \le \sup_{V'\in\mathcal{V}_\epsilon} 2\cdot \bigg\| \sum_{\tau=1}^{t} \phi_\tau \big\{ V'(x_\tau) - \mathbb{E}[V'(x_\tau)|\mathcal{F}_{\tau-1}] \big\} \bigg\|^2_{\Lambda_t^{-1}} + 8t^2\epsilon^2/\lambda. \qquad \text{(E.48)}
\end{aligned}
$$

Now we fixed a $V'\in\mathcal{V}_\epsilon$ and define $\varepsilon_t\in\mathbb{R}^t$ by letting $[\varepsilon_t]_\tau = V'(x_\tau) - \mathbb{E}[V'(x_\tau)|\mathcal{F}_{\tau-1}]$ for any $\tau\ge 1$. Moreover, we define an operator $\Phi\colon \mathcal{H}\to\mathbb{R}^t$ as $\Phi = \big[\phi_1^\top,\ldots,\phi_t^\top\big]^\top$ and let $K_t = \Phi_t\Phi_t^\top \in \mathbb{R}^{t\times t}$. Using this notation, we have $\Lambda_t = \lambda\cdot I_{\mathcal{H}} + \Phi_t^\top\Phi_t$ and

$$
\begin{aligned}
&\bigg\| \sum_{\tau=1}^{t} \phi_\tau \big\{ V'(x_\tau) - \mathbb{E}[V'(x_\tau)|\mathcal{F}_{\tau-1}] \big\} \bigg\|^2_{\Lambda_t^{-1}} = \|\Phi_t^\top \varepsilon_t\|^2_{\Lambda_t^{-1}} = \varepsilon_t^\top \Phi_t \Lambda_t^{-1}\Phi_t^\top \varepsilon_t \\
&\qquad = \varepsilon_t^\top \Phi_t\Phi_t^\top (K_t + \lambda\cdot I)^{-1}\varepsilon_t = \varepsilon_t^\top K_t (K_t+\lambda\cdot I)^{-1}\varepsilon_t, \qquad \text{(E.49)}
\end{aligned}
$$

where the third inequality follows from (C.14). Setting $\lambda = 1+\eta$ for some $\eta>0$, we have

$$
(K_t+\eta\cdot I)\big[K_t + (1+\eta)\cdot I\big]^{-1} = (K_t+\eta\cdot I)\big[I + (K_t+\eta\cdot I)\big]^{-1} = \big[(K_t+\eta\cdot I)^{-1} + I\big]^{-1},
$$

which implies that

$$
\begin{aligned}
\varepsilon_t^\top K_t (K_t+\lambda\cdot I)^{-1}\varepsilon_t &\le \varepsilon_t^\top (K_t+\eta\cdot I)\big[I+(K_t+\eta\cdot I)\big]^{-1}\varepsilon_t \\
&= \varepsilon_t^\top \big[(K_t+\eta\cdot I)^{-1} + I\big]^{-1}\varepsilon_t. \qquad \text{(E.50)}
\end{aligned}
$$

Notice that each entry of ε_t is bounded by H in absolute value since V' is bounded in $[0,H]$. By combining (E.48), (E.49), (E.50), Lemma E.1, and taking a union bound over $\mathcal{V}_\epsilon$, for any $\delta\in(0,1)$, we obtain that, with probability at least $1-\delta$,

$$
\begin{aligned}
&\sup_{V'\in\mathcal{V}_\epsilon} \bigg\| \sum_{\tau=1}^{t} \phi_\tau \big\{ V'(x_\tau) - \mathbb{E}[V'(x_\tau)|\mathcal{F}_{\tau-1}] \big\} \bigg\|^2_{\Lambda_t^{-1}} \\
&\qquad \le H^2\cdot \mathrm{logdet}[(1+\eta)\cdot I + K_t] + 2H^2\cdot \log(\mathcal{N}_\epsilon/\delta) \qquad \text{(E.51)}
\end{aligned}
$$

holds simultaneously for all $t\ge 1$. Besides, notice that $(1+\eta)\cdot I + K_t = [I + (1+\eta)^{-1}\cdot K_t]\cdot[(1+\eta)\cdot I]$, which implies that

$$
\begin{aligned}
\mathrm{logdet}\big[(1+\eta)\cdot I + K_t\big] &= \mathrm{logdet}\big[I + (1+\eta)^{-1}\cdot K_t\big] + t\ln(1+\eta) \\
&\le \mathrm{logdet}\big[I + (1+\eta)^{-1}\cdot K_t\big] + \eta t. \qquad \text{(E.52)}
\end{aligned}
$$

Finally, combining (E.48), (E.51), and (E.52), we conclude that, simultaneously for all $t\ge 1$, (E.45) holds with probability at least $1-\delta$, which concludes the proof. $\square$

Lemma E.3 (Abbasi-Yadkori et al. (2011)). Let $\{\phi_t\}_{t\ge 1}$ be a sequence in the RKHS $\mathcal{H}$. Let $\Lambda_0\colon \mathcal{H}\to\mathcal{H}$ be defined as $\lambda\cdot\mathcal{I}_{\mathcal{H}}$ where $\lambda\ge 1$ and $\mathcal{I}_{\mathcal{H}}$ is the identity mapping on $\mathcal{H}$. For any $t\ge 1$, we define a self-adjoint and positive definite operator Λ_t by letting $\Lambda_t = \Lambda_0 + \sum_{j=1}^{t} \phi_j \phi_j^\top$. Then, for any $t \geq 1$, we have

$$\sum_{j=1}^{t} \min\big\{1, \phi_j^\top \Lambda_{j-1}^{-1} \phi_j\big\} \leq 2\text{logdet}(I + K_t/\lambda),$$

where $K_t \in \mathbb{R}^{t\times t}$ is the Gram matrix obtained from $\{\phi_j\}_{j\in[t]}$, i.e., for any $j, j' \in [t]$, the (j, j')-th entry of K_t is $\langle \phi_j, \phi_{j'} \rangle_{\mathcal{H}}$. Moreover, if we further have $\sup_{t\geq 0} \|\phi_t\|_{\mathcal{H}} \leq 1$, then it holds that

$$\text{logdet}(I + K_t/\lambda) \leq \sum_{j=1}^{t} \phi_j^\top \Lambda_{j-1}^{-1} \phi_j \leq 2\text{logdet}(I + K_t/\lambda).$$

Proof. Note that we have $\log(1 + x) \leq x \leq 2\log(1 + x)$ for all $x \in [0, 1]$. Since Λ_t^{-1} is a self-adjoint and positive definite operator, this implies that

$$\sum_{j=1}^{t} \min\big\{1, \phi_j^\top \Lambda_{j-1}^{-1} \phi_j\big\} \leq \sum_{j=1}^{t} 2\log\big(\min\{2, 1 + \phi_j^\top \Lambda_{j-1}^{-1} \phi_j\}\big) \leq 2\sum_{j=1}^{t} \log\big(1 + \phi_j^\top \Lambda_{j-1}^{-1} \phi_j\big). \tag{E.53}$$

Besides, when additionally it holds that $\sup_{j\geq 1} \|\phi_j\|_{\mathcal{H}} \leq 1$ for all $j \geq 0$, we have

$$\phi_j^\top \Lambda_{j-1}^{-1} \phi_j = \langle \phi_j, \Lambda_{j-1}^{-1} \phi_j \rangle_{\mathcal{H}} \leq \|\phi_j\|_{\mathcal{H}} \cdot \big\|\Lambda_{j-1}^{-1} \phi_j\big\|_{\mathcal{H}} \leq [\lambda_{\min}(\Lambda_0)]^{-1} \cdot \|\phi_j\|_{\mathcal{H}}^2 \leq 1. \tag{E.54}$$

Hence, applying the basic inequality $\log(1 + x) \leq x \leq 2\log(1 + x)$ to (E.54), we have

$$\sum_{j=1}^{t} \log\big(1 + \phi_j^\top \Lambda_{j-1}^{-1} \phi_j\big) \leq \sum_{j=1}^{t} \phi_j^\top \Lambda_{j-1}^{-1} \phi_j \leq 2\sum_{j=1}^{t} \log\big(1 + \phi_j^\top \Lambda_{j-1}^{-1} \phi_j\big). \tag{E.55}$$

For any $j \geq 1$, let $\Lambda_{j-1}^{1/2} \colon \mathcal{H} \to \mathcal{H}$ be the self-adjoint and positive definite operator that is the square-root operator of Λ_{j-1}. Specifically, let $\{\sigma_\ell\}_{\ell\geq 1}$ be the eigevalues of Λ_{j-1} and let $\{v_\ell\}_{\ell\geq 1}$ be the corresponding eigenfunctions. Then $\Lambda_{j-1}^{1/2} = \sum_{\ell\geq 1} \sigma_\ell^{1/2} \cdot v_\ell v_\ell^\top$. Using this notation, for any $j \geq 1$, by the definition of Λ_j, we have

$$\Lambda_j = \Lambda_{j-1} + \phi_j \phi_j^\top = \Lambda_{j-1}^{1/2} \big(\mathcal{I}_{\mathcal{H}} + \Lambda_{j-1}^{-1/2} \phi_j \phi_j^\top \Lambda_{j-1}^{-1/2}\big) \Lambda_{j-1}^{1/2},$$

which implies that

$$\begin{aligned}\text{logdet}(\Lambda_j) &= \text{logdet}(\Lambda_{j-1}) + \text{logdet}\big(\mathcal{I}_{\mathcal{H}} + \Lambda_{j-1}^{-1/2} \phi_j \phi_j^\top \Lambda_{j-1}^{-1/2}\big) \\ &= \text{logdet}\big(\Lambda_{j-1}\big) + \text{logdet}\big(1 + \phi_j^\top \Lambda_{j-1}^{-1} \phi_j\big)\end{aligned} \tag{E.56}$$

Moreover, by direct computation, for any $t \geq 1$, we have

$$\det(\Lambda_t \Lambda_0^{-1}) = \det(I + K_t/\lambda). \tag{E.57}$$

Hence, combining (E.56), and (E.57), we obtain that

$$\sum_{j=1}^{t} \log\big(1 + \phi_j^\top \Lambda_{j-1}^{-1} \phi_j\big) = \text{logdet}(\Lambda_t \Lambda_0^{-1}) = \text{logdet}(I + K_t/\lambda). \tag{E.58}$$

Finally, combining (E.53), (E.55) and (E.58), we conclude the proof of this lemma. □

References

Abbasi-Yadkori, Y., Pál, D. and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*.

Allen-Zhu, Z., Li, Y. and Liang, Y. (2018a). Learning and generalization in overparameterized neural networks, going beyond two layers. *arXiv preprint arXiv:1811.04918*.

Allen-Zhu, Z., Li, Y. and Song, Z. (2018b). A convergence theory for deep learning via over-parameterization. *arXiv preprint arXiv:1811.03962*.

Arora, S., Du, S. S., Hu, W., Li, Z. and Wang, R. (2019). Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *arXiv preprint arXiv:1901.08584*.

Ayoub, A., Jia, Z., Szepesvari, C., Wang, M. and Yang, L. F. (2020). Model-based reinforcement learning with value-targeted regression. *arXiv preprint arXiv:2006.01107*.

Azar, M. G., Osband, I. and Munos, R. (2017). Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*.

Azuma, K. (1967). Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, **19** 357–367.

Bach, F. (2017). Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, **18** 629–681.

Bai, Y. and Lee, J. D. (2019). Beyond linearization: On quadratic and higher-order approximation of wide neural networks. *arXiv preprint arXiv:1910.01619*.

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

Bubeck, S. and Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, **5** 1–122.

Cai, Q., Yang, Z., Jin, C. and Wang, Z. (2019a). Provably efficient exploration in policy optimization. *arXiv preprint arXiv:1912.05830*.

Cai, Q., Yang, Z., Lee, J. D. and Wang, Z. (2019b). Neural temporal-difference learning converges to global optima. In *Advances in Neural Information Processing Systems*.

Calandriello, D., Carratino, L., Lazaric, A., Valko, M. and Rosasco, L. (2019). Gaussian process optimization with adaptive sketching: Scalable and no regret. *arXiv preprint arXiv:1903.05594*.

Cao, Y. and Gu, Q. (2019a). Generalization bounds of stochastic gradient descent for wide and deep neural networks. *arXiv preprint arXiv:1905.13210*.

Cao, Y. and Gu, Q. (2019b). A generalization theory of gradient descent for learning over-parameterized deep ReLU networks. *arXiv preprint arXiv:1902.01384*.

Chizat, L. and Bach, F. (2018). A note on lazy training in supervised differentiable programming. *arXiv preprint arXiv:1812.07956*.

Chowdhury, S. R. and Gopalan, A. (2017). On kernelized multi-armed bandits. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org.

Daniely, A. (2017). SGD learns the conjugate kernel class of the network. In *Advances in Neural Information Processing Systems*.

Dann, C., Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J. and Schapire, R. E. (2018). On oracle-efficient pac rl with rich observations. In *Advances in Neural Information Processing Systems*.

Dann, C., Lattimore, T. and Brunskill, E. (2017). Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*.

Dong, K., Peng, J., Wang, Y. and Zhou, Y. (2019). $\sqrt{n}$-regret for learning in markov decision processes with function approximation and low bellman rank. *arXiv preprint arXiv:1909.02506*.

Du, S. S., Kakade, S. M., Wang, R. and Yang, L. F. (2019a). Is a good representation sufficient for sample efficient reinforcement learning? *arXiv preprint arXiv:1910.03016*.

Du, S. S., Krishnamurthy, A., Jiang, N., Agarwal, A., Dudík, M. and Langford, J. (2019b). Provably efficient RL with rich observations via latent state decoding. *arXiv preprint arXiv:1901.09018*.

Du, S. S., Lee, J. D., Li, H., Wang, L. and Zhai, X. (2018a). Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804*.

Du, S. S., Zhai, X., Poczos, B. and Singh, A. (2018b). Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*.

Duan, Y., Chen, X., Houthooft, R., Schulman, J. and Abbeel, P. (2016). Benchmarking deep reinforcement learning for continuous control. In *International Conference on Machine Learning*.

Durand, A., Maillard, O.-A. and Pineau, J. (2018). Streaming kernel regression with provably adaptive mean, variance, and regularization. *The Journal of Machine Learning Research*, **19** 650–683.

Efroni, Y., Shani, L., Rosenberg, A. and Mannor, S. (2020). Optimistic policy optimization with bandit feedback. *arXiv preprint arXiv:2002.08243*.

Gao, R., Cai, T., Li, H., Hsieh, C.-J., Wang, L. and Lee, J. D. (2019). Convergence of adversarial training in overparametrized neural networks. In *Advances in Neural Information Processing Systems*.

Hofmann, T., Schölkopf, B. and Smola, A. J. (2008). Kernel methods in machine learning. *The annals of statistics* 1171–1220.

Jacot, A., Gabriel, F. and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*.

Jaksch, T., Ortner, R. and Auer, P. (2010). Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, **11** 1563–1600.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J. and Schapire, R. E. (2017). Contextual decision processes with low bellman rank are pac-learnable. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org.

Jin, C., Allen-Zhu, Z., Bubeck, S. and Jordan, M. I. (2018). Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*.

Jin, C., Yang, Z., Wang, Z. and Jordan, M. I. (2019). Provably efficient reinforcement learning with linear function approximation. *arXiv preprint arXiv:1907.05388*.

Kakade, S., Krishnamurthy, A., Lowrey, K., Ohnishi, M. and Sun, W. (2020). Information theoretic regret bounds for online nonlinear control. *arXiv preprint arXiv:2006.12466*.

Krause, A. and Ong, C. S. (2011). Contextual gaussian process bandit optimization. In *Advances in neural information processing systems*.

Krishnamurthy, A., Agarwal, A. and Langford, J. (2016). Pac reinforcement learning with rich observations. In *Advances in Neural Information Processing Systems*.

Lafferty, J. and Lebanon, G. (2005). Diffusion kernels on statistical manifolds. *Journal of Machine Learning Research*, **6** 129–163.

Lattimore, T. and Szepesvári, C. (2018). Bandit algorithms. *preprint*.

Lattimore, T. and Szepesvari, C. (2019). Learning with good feature representations in bandits and in rl with a generative model. *arXiv preprint arXiv:1911.07676*.

Lee, J., Xiao, L., Schoenholz, S. S., Bahri, Y., Sohl-Dickstein, J. and Pennington, J. (2019). Wide neural networks of any depth evolve as linear models under gradient descent. *arXiv preprint arXiv:1902.06720*.

Li, Y. and Liang, Y. (2018). Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*.

Lu, J., Cheng, G. and Liu, H. (2016). Nonparametric heterogeneity testing for massive data. *arXiv preprint arXiv:1601.06212*.

Mendelson, S., Neeman, J. et al. (2010). Regularization in kernel learning. *The Annals of Statistics*, **38** 526–565.

Minh, H. Q., Niyogi, P. and Yao, Y. (2006). Mercer's theorem, feature maps, and smoothing. In *International Conference on Computational Learning Theory*. Springer.

Müller, C. (2012). *Analysis of spherical symmetries in Euclidean spaces*, vol. 129. Springer.

Neyshabur, B. and Li, Z. (2019). Towards understanding the role of over-parametrization in generalization of neural networks. In *International Conference on Learning Representations (ICLR)*.

Osband, I., Aslanides, J. and Cassirer, A. (2018). Randomized prior functions for deep reinforcement learning. In *Advances in Neural Information Processing Systems*.

Osband, I. and Van Roy, B. (2014). Model-based reinforcement learning and the eluder dimension. In *Advances in Neural Information Processing Systems*.

Osband, I., Van Roy, B. and Wen, Z. (2014). Generalization and exploration via randomized value functions. *arXiv preprint arXiv:1402.0635*.

Puterman, M. L. (2014). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.

Rahimi, A. and Recht, B. (2008). Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*.

Rasmussen, C. E. (2003). Gaussian processes in machine learning. In *Summer School on Machine Learning*. Springer.

Russo, D. (2019). Worst-case regret bounds for exploration via randomized value functions. In *Advances in Neural Information Processing Systems*.

Russo, D. and Van Roy, B. (2013). Eluder dimension and the sample complexity of optimistic exploration. In *Advances in Neural Information Processing Systems*.

Sessa, P. G., Bogunovic, I., Kamgarpour, M. and Krause, A. (2019). No-regret learning in unknown games with correlated payoffs. In *Advances in Neural Information Processing Systems*.

Shang, Z., Cheng, G. et al. (2013). Local and global asymptotic inference in smoothing spline models. *Annals of Statistics*, **41** 2608–2638.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M. et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, **529** 484–489.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A. et al. (2017). Mastering the game of Go without human knowledge. *Nature*, **550** 354.

Srinivas, N., Krause, A., Kakade, S. M. and Seeger, M. (2009). Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*.

Srinivas, N., Krause, A., Kakade, S. M. and Seeger, M. W. (2012). Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory*, **58** 3250–3265.

Steinwart, I. and Christmann, A. (2008). *Support vector machines*. Springer Science & Business Media.

Strehl, A. L., Li, L., Wiewiora, E., Langford, J. and Littman, M. L. (2006). PAC model-free reinforcement learning. In *International Conference on Machine Learning*.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Valko, M., Korda, N., Munos, R., Flaounas, I. and Cristianini, N. (2013). Finite-time analysis of kernelised contextual bandits. *arXiv preprint arXiv:1309.6869*.

Van Roy, B. and Dong, S. (2019). Comments on the du-kakade-wang-yang lower bounds. *arXiv preprint arXiv:1911.07910*.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, vol. 47. Cambridge university press.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P. et al. (2019). Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, **575** 350–354.

Wang, R., Salakhutdinov, R. and Yang, L. F. (2020). Provably efficient reinforcement learning with general value function approximation. *arXiv preprint arXiv:2005.10804*.

Wang, W. Y., Li, J. and He, X. (2018). Deep reinforcement learning for nlp. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*.

Wang, Y., Wang, R., Du, S. S. and Krishnamurthy, A. (2019). Optimism in reinforcement learning with generalized linear function approximation. *arXiv preprint arXiv:1912.04136*.

Wen, Z. and Van Roy, B. (2013). Efficient exploration and value function generalization in deterministic systems. In *Advances in Neural Information Processing Systems*.

Wen, Z. and Van Roy, B. (2017). Efficient reinforcement learning in deterministic systems with value function generalization. *Mathematics of Operations Research*, **42** 762–782.

Wu, L., Ma, C. and Weinan, E. (2018). How SGD selects the global minima in over-parameterized learning: A dynamical stability perspective. In *Advances in Neural Information Processing Systems*.

Yang, L. and Wang, M. (2019a). Sample-optimal parametric Q-learning using linearly additive features. In *International Conference on Machine Learning*.

Yang, L. F. and Wang, M. (2019b). Reinforcement leaning in feature space: Matrix bandit, kernels, and regret bound. *arXiv preprint arXiv:1905.10389*.

Yang, Y., Bhattacharya, A. and Pati, D. (2017). Frequentist coverage and sup-norm convergence rate in gaussian process regression. *arXiv preprint arXiv:1708.04753*.

Zanette, A., Brandfonbrener, D., Brunskill, E., Pirotta, M. and Lazaric, A. (2020a). Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*.

Zanette, A., Lazaric, A., Kochenderfer, M. and Brunskill, E. (2020b). Learning near optimal policies with low inherent bellman error. *arXiv preprint arXiv:2003.00153*.

Zhang, Y., Duchi, J. and Wainwright, M. (2015). Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *Journal of Machine Learning Research*, **16** 3299–3340.

Zhou, D., He, J. and Gu, Q. (2020). Provably efficient reinforcement learning for discounted mdps with feature mapping. *arXiv preprint arXiv:2006.13165*.

Zhou, D., Li, L. and Gu, Q. (2019). Neural contextual bandits with upper confidence bound-based exploration. *arXiv preprint arXiv:1911.04462*.

Zou, D., Cao, Y., Zhou, D. and Gu, Q. (2018). Stochastic gradient descent optimizes overparameterized deep ReLU networks. *arXiv preprint arXiv:1811.08888*.