On Function Approximation in Reinforcement Learning: Optimism in the Face of Large State Spaces

Zhuoran Yang
Princeton University
zy6@princeton.edu

Chi Jin
Princeton University
chij@princeton.edu

Zhaoran Wang
Northwestern University
zhaoran.wang@northwestern.edu

Mengdi Wang
Princeton University
mengdiw@princeton.edu

Michael I. Jordan
University of California, Berkeley
jordan@cs.berkeley.edu

Abstract

The classical theory of reinforcement learning (RL) has focused on tabular and linear representations of value functions. Further progress hinges on combining RL with modern function approximators such as kernel functions and deep neural networks, and indeed there have been many empirical successes that have exploited such combinations in large-scale applications. There are profound challenges, however, in developing a theory to support this enterprise, most notably the need to take into consideration the exploration-exploitation tradeoff at the core of RL in conjunction with the computational and statistical tradeoffs that arise in modern function-approximation-based learning systems. We approach these challenges by studying an optimistic modification of the least-squares value iteration algorithm, in the context of the action-value function represented by a kernel function or an overparameterized neural network. We establish both polynomial runtime complexity and polynomial sample complexity for this algorithm, without additional assumptions on the data-generating model. In particular, we prove that the 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, a result which exhibits clearly the benefit of function approximation in RL.

1 Introduction

Reinforcement learning (RL) algorithms combined with modern function approximators such as kernel functions and deep neural networks have produced empirical successes in a variety of application problems [e.g., 27, 60, 61, 72, 70]. However, theory has lagged, and when these powerful function approximators are employed, there is little theoretical guidance regarding the design of RL algorithms that are efficient computationally or statistically, or regarding whether they even converge. In particular, function approximation blends statistical estimation issues with dynamical optimization issues, resulting in the need to balance the bias-variance tradeoffs that arise in statistical estimation with the exploration-exploitation tradeoffs that are inherent in RL. Accordingly, full theoretical treatments are mostly restricted to the tabular setting, where both the state and action spaces are

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

discrete and the value function can be represented as a table [see, e.g., 33, 52, 6, 35, 50, 56], and there is a disconnect between theory and the most compelling applications.

Provably efficient exploration in the function approximation setting has been addressed only recently, with most of the existing work considering (generalized) linear models [78, 77, 36, 12, 80, 73]. These algorithms and their analyses stem from classical upper confidence bound (UCB) or Thompson sampling methods for linear contextual bandits [11, 41] and it seems difficult to extend them beyond the linear setting. Unfortunately, the linear assumption is rather rigid and rarely satisfied in practice; moreover, when such a model is misspecified, sublinear regret guarantees can vanish. There has been some recent work that has presented sample-efficient algorithms with general function approximation. However, these methods are either computationally intractable [39, 34, 20, 22] or hinge on strong assumptions on the transition model [75, 24]. Thus, the following question remains open:

Can we design RL algorithms that incorporate powerful nonlinear function approximators such as neural networks or kernel functions and provably achieve both computational and statistical efficiency?

In this work, we provide an affirmative answer to this question. Focusing on the setting of an episodic Markov decision process (MDP) where the value function is represented by either a kernel function or an overparameterized neural network, we propose an RL algorithm with polynomial runtime complexity and sample complexity, without imposing any additional assumptions on the data-generating model. Our algorithm is relatively simple—it is an optimistic modification of the least-squares value iteration algorithm (LSVI) [10]—a classical batch RL algorithm—to which we add a UCB bonus term to each iterate. Specifically, when using a kernel function, each LSVI update becomes a kernel ridge regression, and the bonus term is derived from that proposed for kernelized contextual bandits [62, 67, 18]. For the neural network setting, motivated by the NeuralUCB algorithm for contextual bandits [84], we construct a UCB bonus from the tangent features of the neural network and we perform the LSVI updates via projected gradient descent. In both of these settings, the usage of the UCB bonus ensures that the value functions constructed by the algorithm are always optimistic in the sense that they serve as uniform 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, where 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 function class used to represent the value function 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 depend 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. This clearly exhibits the benefit of function approximation by tying it directly to sample efficiency. To the best of our knowledge, this is the first provably efficient framework for reinforcement learning with kernel and neural network function approximations.

Related Work. There is a vast literature on establishing provably efficient RL methods in the absence of a generative model or an explorative behavioral policy. Much of this literature has focused on the tabular setting; see [33, 52, 6, 21, 65, 35, 56] and the references therein. In particular, [6, 35] prove that an 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, algorithms designed for the tabular setting cannot be directly applied to the function approximation setting, where the number of effective states is large. A recent literature has accordingly focused on the function approximation setting, specifically the (generalized) linear setting where the value function (or the transition model) can be represented using a linear transform of a known feature mapping [77, 78, 36, 12, 80, 73, 5, 83, 37]. Among these papers, our work is most closely related to [36]. In particular, in our kernel setting when the kernel function has a finite rank, both our LSVI algorithm and the corresponding regret bound reduce to those established in [36]. However, the sample complexity and regret bounds in [36] diverge when the dimension of the feature mapping goes to infinity and thus cannot be directly applied to the kernel setting.

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	$H^2 \cdot \sqrt{(\log T)^{3/\gamma} \cdot T} \cdot \log(TH)$
overparameterized neural network	$H^2 \cdot \sqrt{d_{\text{eff}} \cdot [d_{\text{eff}} + \log N_\infty(\epsilon^*)]} \cdot \sqrt{T} + \text{poly}(T, H) \cdot m^{-1/12}$

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 $\epsilon^* = H/T$. Note that to obtain concrete bounds, we apply the general result to RKHS's with various eigenvalue decay conditions. Here γ is a positive integer in the case of γ-finite spectrum and is a positive number in the case of γ-exponential decay. 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 $\text{poly}(T, H)$ is a polynomial in T and H. Such a general regret bound can be expressed concretely as a function of the spectrum of the NTK.

Also closely related to our work is [71], which studies a similar optimistic LSVI algorithm for general function approximation. This work focuses on value function classes with bounded eluder dimension [57, 51]. It is unclear whether whether this formulation can be extended to the kernel or neural network settings. [78] studies a kernelized MDP model where the transition model can be directly estimated. Under a slightly more general model, [5] recently propose an optimistic model-based algorithm via value-targeted regression, where the model class is the set of functions with bounded eluder dimension. In other recent work, [37] studies a nonlinear control formulation in which the transition dynamics belongs to a known RKHS and can be directly estimated from the data. Our work differs from this work in that we impose an explicit assumption on the transition model and our proposed algorithm is model-free.

Other authors who have presented regret bounds and sample complexities beyond the linear setting include [39, 34, 20, 22]. These algorithms generally involve either high computational costs or require possibly restrictive assumptions on the transition model [74, 75, 24].

Our work is also related to the literature on contextual bandits with either kernel function classes [62, 38, 63, 67, 18, 28] or neural network function classes [84]. Our construction of a bonus function for the RL setting has been adopted from this previous work. However, while contextual bandits can be viewed formally as special cases of our episodic MDP formulation with the episode length equal to one, the temporal dependence in the MDP setting raises significant challenges. In particular, the covering number $N_\infty(\epsilon^*)$ in Table 1 arises as a consequence of the fundamental challenge of performing temporally extended exploration in RL.

Finally, our analysis of the optimistic LSVI algorithm is related to recent work on optimization and generalization in overparameterized neural networks within the framework of the neural tangent kernel [32]. See also [19, 32, 76, 25, 26, 3, 2, 85, 17, 44, 4, 15, 16, 43]. This literature focuses principally on supervised learning, however; in the RL setting we need an additional bonus term in the least-squares problem and thus require a novel analysis.

2 Background

In this section, we provide essential background on reinforcement learning, reproducing kernel Hilbert space (RKHS), and overparameterized neural networks.

Episodic Markov Decision Processes

We focus on episodic MDPs, denoted $\text{MDP}(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, respectively, the integer $H > 0$ is the length of each episode, $\mathbb{P} = \{\mathbb{P}_h\}_{h\in[H]}$ and $r = \{r_h\}_{h\in[H]}$ are the Markov transition kernel and the reward functions, respectively, where we let $[n]$

denote the set $\{1, \dots, n\}$ for integers $n \geq 1$. We assume that $\mathcal{S}$ is a measurable space of possibly infinite cardinality while $\mathcal{A}$ is a finite set. Finally, for each $h \in [H]$, $\mathbb{P}_h(\cdot \,|\, x, a)$ denotes the probability transition kernel when 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 the probability of the agent taking action a at state x at the h-th step in the episode.

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 $r_h(x_h^t, a_h^t)$. The environment then transitions 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 thus the final reward that the agent receives.

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

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

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

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

Moreover, let $\pi^\star$ denote the optimal policy which by definition 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 write

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

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 [11] of the agent is defined as

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

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

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 [53, 66] 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

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

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 [10, 52] 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

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

where $\mathrm{pen}(f)$ is a regularization term. Moreover, to foster exploration, following the principle of optimism in the face of uncertainty [66], 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 [36], which further reduces to the tabular UCBVI algorithm [6] 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 [62, 67, 18, 28, 78, 58, 14]. With these two modifications, we obtain the Kernel Optimistic Least-Squares Value Iteration (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

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

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

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

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 $\widehat{Q}_h^t$ [55]. Such a bonus term also appears in the literature on kernelized contextual bandits [62, 67, 18, 28, 78, 58, 14] and is reduced to the UCB bonus proposed for linear bandits [11, 41] when the feature mapping ϕ of the RKHS is finite-dimensional. In this case, KOVI reduces to the LSVI-UCB algorithm proposed in [36] 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.

4 Theory of Kernel Optimistic Least-Squares Value Iteration

In this section, we prove that KOVI achieves $\mathcal{O}(\delta_{\mathcal{H}} H^2 \sqrt{T})$-regret bounds, where $\delta_{\mathcal{H}}$ characterizes the intrinsic complexity of the RKHS $\mathcal{H}$ that is used to approximate $\{Q_h^\star\}_{h\in[H]}$. 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 $\mathcal{Q}^\star = \{f \in \mathcal{H}\colon \|f\|_{\mathcal{H}} \le R_Q H\}$. We assume that for any $h \in [H]$ and any $Q\colon \mathcal{S} \times \mathcal{A} \to [0, H]$, we have $\mathbb{T}_h^\star Q \in \mathcal{Q}^\star$.

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 [23] 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

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

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 $V\colon \mathcal{S} \to [0, H]$, 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 [77, 78, 36, 80, 81, 73] for (generalized) linear functions. Also see [23, 68, 42] 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 [62], 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

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

As we will show in Lemma H.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

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

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

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

As shown in (D.6), 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 (D.6) to

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

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 $N_\infty(\epsilon^*, h, B_T) \le |\mathcal{S}\times\mathcal{A}|\cdot\log T$. 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 (D.6) holds for any RKHS in general. It hinges on (i) Assumption 4.1, which postulates that the RKHS-norm ball $\{f\in\mathcal{H}\colon \|f\|_{\mathcal{H}} \le R_Q H\}$ 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 [62, 18]. 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 (D.6), 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 (B.1) has eigenvalues $\{\sigma_j\}_{j\ge1}$ and eigenfunctions $\{\psi_j\}_{j\ge1}$. We assume that $\{\sigma_j\}_{j\ge1}$ satisfies one of the following two 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 $\sigma_j \le C_1\cdot\exp(-C_2\cdot j^\gamma)$ for all $j \ge 1$.

Moreover, for case (ii), we further assume that there exist constants $\tau\in[0,1/2)$ $C_\psi > 0$ such that $\sup_{z\in\mathcal{Z}} \sigma_j^\tau\cdot|\psi_j(z)| \le C_\psi$ for all $j\ge1$.

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 [77, 78, 36, 80]. Moreover, when $\mathcal{H}$ satisfies case (i), KOVI reduces to the LSVI-UCB algorithm studied in [36]. In addition, case (ii) postulates that the eigenvalues of T_K decays exponentially fast, 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 $\gamma = 1/d$ [62]. Moreover, we assume that there exists $\tau\in[0,1/2)$ such that $\sigma_j^\tau\cdot\|\psi_j\|_\infty$ is universally bounded. Since $K(z,z)\le1$, 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 nonparametric statistics [40, 59, 82, 45, 79]. It can be shown that the squared exponential kernel on unit sphere in $\mathbb{R}^d$ satisfy this condition for any $\tau > 0$. See [46] for a more detailed discussion.

Now we present the regret bounds for the two 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

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

Here C_b is an absolute constant that does not depend on T or H. Then, there exists an absolute constant C_r such that, with probability at least $1 - (T^2H^2)^{-1}$, we have

$$\mathrm{Regret}(T) \le \begin{cases} C_r \cdot H^2 \cdot \sqrt{\gamma^3 T} \cdot \log(\gamma TH) & \gamma\text{-finite spectrum,} \\ C_r \cdot H^2 \cdot \sqrt{(\log T)^{3/\gamma} \cdot T} \cdot \log(TH) & \gamma\text{-exponential decay.} \end{cases} \tag{4.9}$$

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 both the two cases specified in Assumption 4.3. Note that the linear MDP [36] satisfies the γ-finite spectrum condition and KOVI recovers the LSVI-UCB algorithm studied in [36] when restricted to this setting. Moreover, our $\widetilde{\mathcal{O}}(H^2 \cdot \sqrt{\gamma^3 T})$ also matches the regret bound in [36]. In addition, under the γ-exponential eigenvalue decay condition, as we will show in §I, 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 (D.6). 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 [62] for kernel contextual bandits, which is due to bounding the log-covering number. See §G.1 for details.

Furthermore, similarly to the discussion in Section 3.1 of [35], the regret bound in (D.6) 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 two cases considered in Assumption 4.3, such a sample complexity guarantee reduces to $\widetilde{\mathcal{O}}\big(H^4 \cdot \gamma^3/\epsilon^2\big)$ and $\widetilde{\mathcal{O}}\big(H^4 \cdot (\log T)^{2+3/\gamma}/\epsilon^2\big)$, respectively. Moreover, similar to [36], 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 §D.

5 Conclusion

In this paper, we have presented an algorithmic framework for reinforcement learning with general function approximation. Such a framework is based on an optimistic least-squares value iteration algorithm that incorporates an additional bonus term in the solution to a least-squares value estimation problem. The bonus term promotes exploration. When deploying this framework in the settings of kernel function and overparameterized neural networks, respectively, we obtain two algorithms KOVI and NOVI. Both algorithms are provably efficient, both computationally and in terms of the number of samples. Specifically, under the kernel and neural network settings respectively, KOVI and NOVI both achieve sublinear regret, $\widetilde{O}(\delta_{\mathcal{F}} H^2 \sqrt{T})$, where $\delta_{\mathcal{F}}$ is a quantity that characterizes the intrinsic complexity of the function class $\mathcal{F}$. To the best of our knowledge, this is the first provably efficient reinforcement learning algorithm in the general settings of kernel and neural function approximations.

Broader Impact

This is a theoretical paper. We do not foresee our work directly having any societal consequences. However, reinforcement learning is a tool that is increasingly used in practical machine learning applications, especially in the setting where nonlinear function approximation is involved. Theoretical explorations related to reinforcement learning with function approximation may help provide frameworks through which to reason about, and design safer and more reliable practical systems.

Acknowledgements

We would like to thank the Simons Institute for the Theory of Computing in Berkeley, where this project was initiated. Zhuoran Yang would like to thank Jianqing Fan, Csaba Szepsvári, Tuo Zhao, Simon Shaolei Du, Ruosong Wang, and Yiping Lu for valuable discussions. 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. Michael Jordan gratefully acknowledges funding from the Mathematical Data Science program of the Office of Naval Research under grant number N00014-18-1-2764.

References

[1] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

[2] Z. Allen-Zhu, Y. Li, and Y. Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *arXiv preprint arXiv:1811.04918*, 2018.

[3] Z. Allen-Zhu, Y. Li, and Z. Song. A convergence theory for deep learning via over-parameterization. *arXiv preprint arXiv:1811.03962*, 2018.

[4] S. Arora, S. S. Du, W. Hu, Z. Li, and R. Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *arXiv preprint arXiv:1901.08584*, 2019.

[5] A. Ayoub, Z. Jia, C. Szepesvari, M. Wang, and L. F. Yang. Model-based reinforcement learning with value-targeted regression. *arXiv preprint arXiv:2006.01107*, 2020.

[6] M. G. Azar, I. Osband, and R. Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272, 2017.

[7] K. Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367, 1967.

[8] F. Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

[9] Y. Bai and J. D. Lee. Beyond linearization: On quadratic and higher-order approximation of wide neural networks. *arXiv preprint arXiv:1910.01619*, 2019.

[10] S. J. Bradtke and A. G. Barto. Linear least-squares algorithms for temporal difference learning. *Machine learning*, 22(1-3):33–57, 1996.

[11] S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

[12] Q. Cai, Z. Yang, C. Jin, and Z. Wang. Provably efficient exploration in policy optimization. *arXiv preprint arXiv:1912.05830*, 2019.

[13] Q. Cai, Z. Yang, J. D. Lee, and Z. Wang. Neural temporal-difference learning converges to global optima. In *Advances in Neural Information Processing Systems*, pages 11312–11322, 2019.

[14] D. Calandriello, L. Carratino, A. Lazaric, M. Valko, and L. Rosasco. Gaussian process optimization with adaptive sketching: Scalable and no regret. *arXiv preprint arXiv:1903.05594*, 2019.

[15] Y. Cao and Q. Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. *arXiv preprint arXiv:1905.13210*, 2019.

[16] Y. Cao and Q. Gu. A generalization theory of gradient descent for learning over-parameterized deep ReLU networks. *arXiv preprint arXiv:1902.01384*, 2019.

[17] L. Chizat and F. Bach. A note on lazy training in supervised differentiable programming. *arXiv preprint arXiv:1812.07956*, 2018.

[18] S. R. Chowdhury and A. Gopalan. On kernelized multi-armed bandits. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 844–853. JMLR. org, 2017.

[19] A. Daniely. SGD learns the conjugate kernel class of the network. In *Advances in Neural Information Processing Systems*, 2017.

[20] C. Dann, N. Jiang, A. Krishnamurthy, A. Agarwal, J. Langford, and R. E. Schapire. On oracle-efficient pac rl with rich observations. In *Advances in Neural Information Processing Systems*, pages 1422–1432, 2018.

[21] C. Dann, T. Lattimore, and E. Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 5713–5723, 2017.

[22] K. Dong, J. Peng, Y. Wang, and Y. Zhou. $\sqrt{n}$-regret for learning in markov decision processes with function approximation and low bellman rank. *arXiv preprint arXiv:1909.02506*, 2019.

[23] S. S. Du, S. M. Kakade, R. Wang, and L. F. Yang. Is a good representation sufficient for sample efficient reinforcement learning? *arXiv preprint arXiv:1910.03016*, 2019.

[24] S. S. Du, A. Krishnamurthy, N. Jiang, A. Agarwal, M. Dudík, and J. Langford. Provably efficient RL with rich observations via latent state decoding. *arXiv preprint arXiv:1901.09018*, 2019.

[25] S. S. Du, J. D. Lee, H. Li, L. Wang, and X. Zhai. Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804*, 2018.

[26] S. S. Du, X. Zhai, B. Poczos, and A. Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.

[27] Y. Duan, X. Chen, R. Houthooft, J. Schulman, and P. Abbeel. Benchmarking deep reinforcement learning for continuous control. In *International Conference on Machine Learning*, pages 1329–1338, 2016.

[28] A. Durand, O.-A. Maillard, and J. Pineau. Streaming kernel regression with provably adaptive mean, variance, and regularization. *The Journal of Machine Learning Research*, 19(1):650–683, 2018.

[29] Y. Efroni, L. Shani, A. Rosenberg, and S. Mannor. Optimistic policy optimization with bandit feedback. *arXiv preprint arXiv:2002.08243*, 2020.

[30] R. Gao, T. Cai, H. Li, C.-J. Hsieh, L. Wang, and J. D. Lee. Convergence of adversarial training in overparametrized neural networks. In *Advances in Neural Information Processing Systems*, pages 13009–13020, 2019.

[31] T. Hofmann, B. Schölkopf, and A. J. Smola. Kernel methods in machine learning. *The annals of statistics*, pages 1171–1220, 2008.

[32] A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, 2018.

[33] T. Jaksch, R. Ortner, and P. Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4):1563–1600, 2010.

[34] N. Jiang, A. Krishnamurthy, A. Agarwal, J. Langford, and R. E. Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1704–1713. JMLR. org, 2017.

[35] C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873, 2018.

[36] C. Jin, Z. Yang, Z. Wang, and M. I. Jordan. Provably efficient reinforcement learning with linear function approximation. *arXiv preprint arXiv:1907.05388*, 2019.

[37] S. Kakade, A. Krishnamurthy, K. Lowrey, M. Ohnishi, and W. Sun. Information theoretic regret bounds for online nonlinear control. *arXiv preprint arXiv:2006.12466*, 2020.

[38] A. Krause and C. S. Ong. Contextual gaussian process bandit optimization. In *Advances in neural information processing systems*, pages 2447–2455, 2011.

[39] A. Krishnamurthy, A. Agarwal, and J. Langford. Pac reinforcement learning with rich observations. In *Advances in Neural Information Processing Systems*, pages 1840–1848, 2016.

[40] J. Lafferty and G. Lebanon. Diffusion kernels on statistical manifolds. *Journal of Machine Learning Research*, 6(Jan):129–163, 2005.

[41] T. Lattimore and C. Szepesvári. Bandit algorithms. *preprint*, 2018.

[42] T. Lattimore and C. Szepesvari. Learning with good feature representations in bandits and in rl with a generative model. *arXiv preprint arXiv:1911.07676*, 2019.

[43] J. Lee, L. Xiao, S. S. Schoenholz, Y. Bahri, J. Sohl-Dickstein, and J. Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *arXiv preprint arXiv:1902.06720*, 2019.

[44] Y. Li and Y. Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, 2018.

[45] J. Lu, G. Cheng, and H. Liu. Nonparametric heterogeneity testing for massive data. *arXiv preprint arXiv:1601.06212*, 2016.

[46] S. Mendelson, J. Neeman, et al. Regularization in kernel learning. *The Annals of Statistics*, 38(1):526–565, 2010.

[47] H. Q. Minh, P. Niyogi, and Y. Yao. Mercer's theorem, feature maps, and smoothing. In *International Conference on Computational Learning Theory*, pages 154–168. Springer, 2006.

[48] C. Müller. *Analysis of spherical symmetries in Euclidean spaces*, volume 129. Springer, 2012.

[49] B. Neyshabur and Z. Li. Towards understanding the role of over-parametrization in generalization of neural networks. In *International Conference on Learning Representations (ICLR)*, 2019.

[50] I. Osband, J. Aslanides, and A. Cassirer. Randomized prior functions for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 8617–8629, 2018.

[51] I. Osband and B. Van Roy. Model-based reinforcement learning and the eluder dimension. In *Advances in Neural Information Processing Systems*, pages 1466–1474, 2014.

[52] I. Osband, B. Van Roy, and Z. Wen. Generalization and exploration via randomized value functions. *arXiv preprint arXiv:1402.0635*, 2014.

[53] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

[54] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pages 1177–1184, 2008.

[55] C. E. Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer, 2003.

[56] D. Russo. Worst-case regret bounds for exploration via randomized value functions. In *Advances in Neural Information Processing Systems*, pages 14410–14420, 2019.

[57] D. Russo and B. Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In *Advances in Neural Information Processing Systems*, pages 2256–2264, 2013.

[58] P. G. Sessa, I. Bogunovic, M. Kamgarpour, and A. Krause. No-regret learning in unknown games with correlated payoffs. In *Advances in Neural Information Processing Systems*, pages 13602–13611, 2019.

[59] Z. Shang, G. Cheng, et al. Local and global asymptotic inference in smoothing spline models. *Annals of Statistics*, 41(5):2608–2638, 2013.

[60] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[61] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354, 2017.

[62] N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

[63] N. Srinivas, A. Krause, S. M. Kakade, and M. W. Seeger. Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory*, 58(5):3250–3265, 2012.

[64] I. Steinwart and A. Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

[65] A. L. Strehl, L. Li, E. Wiewiora, J. Langford, and M. L. Littman. PAC model-free reinforcement learning. In *International Conference on Machine Learning*, pages 881–888, 2006.

[66] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[67] M. Valko, N. Korda, R. Munos, I. Flaounas, and N. Cristianini. Finite-time analysis of kernelised contextual bandits. *arXiv preprint arXiv:1309.6869*, 2013.

[68] B. Van Roy and S. Dong. Comments on the du-kakade-wang-yang lower bounds. *arXiv preprint arXiv:1911.07910*, 2019.

[69] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[70] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

[71] R. Wang, R. Salakhutdinov, and L. F. Yang. Provably efficient reinforcement learning with general value function approximation. *arXiv preprint arXiv:2005.10804*, 2020.

[72] W. Y. Wang, J. Li, and X. He. Deep reinforcement learning for nlp. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 19–21, 2018.

[73] Y. Wang, R. Wang, S. S. Du, and A. Krishnamurthy. Optimism in reinforcement learning with generalized linear function approximation. *arXiv preprint arXiv:1912.04136*, 2019.

[74] Z. Wen and B. Van Roy. Efficient exploration and value function generalization in deterministic systems. In *Advances in Neural Information Processing Systems*, pages 3021–3029, 2013.

[75] Z. Wen and B. Van Roy. Efficient reinforcement learning in deterministic systems with value function generalization. *Mathematics of Operations Research*, 42(3):762–782, 2017.

[76] L. Wu, C. Ma, and E. Weinan. How SGD selects the global minima in over-parameterized learning: A dynamical stability perspective. In *Advances in Neural Information Processing Systems*, 2018.

[77] L. Yang and M. Wang. Sample-optimal parametric Q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004, 2019.

[78] L. F. Yang and M. Wang. Reinforcement leaning in feature space: Matrix bandit, kernels, and regret bound. *arXiv preprint arXiv:1905.10389*, 2019.

[79] Y. Yang, A. Bhattacharya, and D. Pati. Frequentist coverage and sup-norm convergence rate in gaussian process regression. *arXiv preprint arXiv:1708.04753*, 2017.

[80] A. Zanette, D. Brandfonbrener, E. Brunskill, M. Pirotta, and A. Lazaric. Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pages 1954–1964, 2020.

[81] A. Zanette, A. Lazaric, M. Kochenderfer, and E. Brunskill. Learning near optimal policies with low inherent bellman error. *arXiv preprint arXiv:2003.00153*, 2020.

[82] Y. Zhang, J. Duchi, and M. Wainwright. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *Journal of Machine Learning Research*, 16(1):3299–3340, 2015.

[83] D. Zhou, J. He, and Q. Gu. Provably efficient reinforcement learning for discounted mdps with feature mapping. *arXiv preprint arXiv:2006.13165*, 2020.

[84] D. Zhou, L. Li, and Q. Gu. Neural contextual bandits with upper confidence bound-based exploration. *arXiv preprint arXiv:1911.04462*, 2019.

[85] D. Zou, Y. Cao, D. Zhou, and Q. Gu. Stochastic gradient descent optimizes over-parameterized deep ReLU networks. *arXiv preprint arXiv:1811.08888*, 2018.

A 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., [33, 52, 6, 21, 65, 35, 56] and the references therein. It is shown in [6, 35] 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, [77, 78, 36, 12, 80, 73, 5, 83, 37] 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 [36]. 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 [36]. 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 [71], which studies a similar optimistic LSVI algorithm for general function approximation. Their work focuses on value function classes with bounded eluder dimensions [57, 51] and it is unclear whether their construction of the bonus function can be extended to the kernel or neural settings. Besides, [78] also study a kernelized MDP model where the transition model can be directly estimated. Under a slightly more general model, [5] 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, [37] 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 [39, 34, 20, 22] or require additional assumptions on the transition model that might be restrictive [74, 75, 24]

In addition, our work is also related to the literature on contextual bandits with kernel or [62, 38, 63, 67, 18, 28] neural network functions [84], 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 [62, 18].

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 [32]. Most of these works focus on the supervised learning [19, 32, 76, 25, 26, 3, 2, 85, 17, 44, 4, 15, 16, 43]. In contrast, our algorithm incorporates an additional bonus term in the least-squares problem and thus requires novel analysis.

B Additional Background

In this section, we present the background of reproducing kernel Hilbert space and overparameterized neural networks.

B.1 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) [31]. 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

$$T_K f(z) = \int_{\mathcal{Z}} K(z, z') \cdot f(z') \, \mathrm{d}z', \qquad \forall f \in \mathcal{L}^2(\mathcal{Z}). \tag{B.1}$$

By Mercer's Theorem [64], 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

$$K(z, z') = \sum_{i=1}^{\infty} \sigma_i \cdot \psi_i(z) \cdot \psi_j(z'). \tag{B.2}$$

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

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

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 $\{\sqrt{\sigma_i}\psi_i\}_{i \ge 1}$ 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 $z \in \mathcal{Z}$.

B.2 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

$$f(z; b, W) = \frac{1}{\sqrt{2m}} \sum_{j=1}^{2m} b_j \cdot \mathrm{act}(W_j^\top z), \qquad \forall z \in \mathcal{Z}. \tag{B.3}$$

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 [30, 9] 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 [49, 2, 4]. Under the such a regime, the dynamics of training neural networks are well captured by the framework of neural tangent kernel (NTK) [32]. 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

$$\varphi(z; W) = \nabla_W f(z; W) = \big(\nabla_{W_1} f(z; W), \ldots, \nabla_{W_{2m}} f(z; W)\big), \qquad \forall z \in \mathcal{Z}. \tag{B.4}$$

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

$$K_m(z, z') = \big\langle \varphi(z; W^{(0)}), \varphi(z'; W^{(0)}) \big\rangle, \qquad \forall (z, z') \in \mathcal{Z} \times \mathcal{Z}. \tag{B.5}$$

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 $W^{(0)}$,

$$f(\cdot; W) \approx \widehat{f}(\cdot; W) = f(\cdot, W^{(0)}) + \big\langle \phi(\cdot; W^{(0)}), W - W^{(0)} \big\rangle = \big\langle \phi(\cdot; W^{(0)}), W - W^{(0)} \big\rangle. \tag{B.6}$$

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 $K_{\mathrm{ntk}}\colon \mathcal{Z} \times \mathcal{Z}$, dubbed as neural tangent kernel (NTK), which is given by

$$K_{\mathrm{ntk}}(z, z') = \mathbb{E}\big[\mathrm{act}'(w^\top z) \cdot \mathrm{act}'(w^\top z') \cdot z^\top z'\big], \qquad (z, z') \in \mathcal{Z} \times \mathcal{Z}, \tag{B.7}$$

where act' is the derivative of the activation function, and the expectation in (B.7) is taken with respect to $w \sim N(0, I_d/d)$.

C Kernel and Neural Optimistic Least-Squares Value Iteration

In this section, we lay out the details of KOVI and NOVI, which are omitted for brevity. We remark that the loss function L_h^t in Line 7 of Algorithm 4 is given in (C.1) and its global minimizer $\widehat{W}_h^t$ can be efficiently obtained by first-order optimization methods.

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 $a_h^t \leftarrow \mathrm{argmax}_{a \in \mathcal{A}} Q_h^t(x_h^t, a)$.
    Observe the reward $r_h(x_h^t, a_h^t)$ and the next state x_{h+1}^t.
  end for
end for

C.1 Neural Optimistic Value Iteration

In this subsection, 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 (B.3), which is initialized via the symmetric initialization scheme [30, 9] introduced in §B.2. Moreover, for simplicity, we assume

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

Input: Parameters λ and β.
Initialize the network weights $(b^{(0)}, W^{(0)})$ via the symmetric initialization scheme.
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**
    Solve the neural network optimization problem $\widehat{W}_h^t = \operatorname{argmin}_W L_h^t(W)$.
    Update $\Lambda_h^t = \Lambda_h^{t-1} + \varphi(x_h^{t-1}, a_h^{t-1}; \widehat{W}_h^t)\varphi(x_h^{t-1}, a_h^{t-1}; \widehat{W}_h^t)^\top$.
    Obtain the bonus function b_h^t defined in (C.4).
    Obtain value functions

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

  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

that all the neural networks share the same initial weights, denoted by $(b^{(0)}, W^{(0)})$. Besides, we fix $b = b^{(0)}$ in (B.3) 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

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

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

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

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 [26, 25, 4]. 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 (B.4). We define matrix $\Lambda_h^t \in \mathbb{R}^{2md\times 2md}$ as

$$\Lambda_h^t = \lambda \cdot I_{2md} + \sum_{\tau=1}^{t-1}\varphi\big(x_h^\tau, a_h^\tau; \widehat{W}_h^{\tau+1}\big)\varphi\big(x_h^\tau, a_h^\tau; \widehat{W}_h^{\tau+1}\big)^\top, \tag{C.3}$$

which can be recursively computed by letting

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

Then the bonus function b_h^t is defined as

$$b_h^t(x,a) = \big[\varphi\big(x,a;\widehat{W}_h^t\big)^\top(\Lambda_h^t)^{-1}\varphi\big(x,a;\widehat{W}_h^t\big)\big]^{1/2}, \qquad \forall (x,a) \in \mathcal{S}\times\mathcal{A}. \tag{C.4}$$

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 (C.2) and (C.4), respectively. By letting π^t be the greedy policy with respect to $\{Q_h^t\}_{h\in[H]}$, we obtain the Neural Optimistic Least-Squares Value Iteration (NOVI) algorithm, whose details are stated in Algorithm 4 in §F.

The intuition of the bonus term in (C.4) 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 (C.3) and (C.4) by $\varphi(\cdot;W^{(0)})$, then b_h^t recovers the UCB bonus in linear contextual bandits and linear MDPs with feature mapping $\varphi(\cdot;W^{(0)})$ [1, 36, 73]. 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.

D Theory of Neural Optimistic Least-Squares Value Iteration

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 (B.7). 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 §B.2. 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

$$\mathcal{Q}^\star=\bigg\{f_\alpha(z)=\int_{\mathbb{R}^d}\text{act}'(w^\top z)\cdot z^\top\alpha(w)\,\mathrm{d}p_0(w)\colon \alpha\colon\mathbb{R}^d\to\mathbb{R}^d,\|\alpha\|_{2,\infty}\le R_Q H/\sqrt{d}\bigg\},\quad\text{(D.1)}$$

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 [30], $\mathcal{Q}^\star$ is a dense subset of the RKHS $\mathcal{H}$. Thus, when R_Q is sufficiently large, $\mathcal{Q}^\star$ in (D.1) is an expressive function class. We impose the following condition on $\mathcal{Q}^\star$.

Assumption D.1. 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 D.1 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 (D.1) are dense in $\mathcal{H}$, Assumptions D.1 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 (C.3), 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 (B.4). With a slight abuse of notation, for any $R,B>0$, we let $\mathcal{Q}_{\text{ucb}}(h,R,B)$ denote that class of functions that take the form of

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

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, $\mathcal{Q}_{\text{ucb}}(h,R,B)$ 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 $\mathcal{Q}_{\text{ucb}}(h,R,B)$ 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 (B.5). 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, $\mathcal{Q}_{\text{ucb}}(h,R,B)$ 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 $\mathcal{Q}_{\text{ucb}}(h,R_T,B)$ with respect to the ℓ_∞-norm on $\mathcal{Z}$.

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

Theorem D.2. Under Assumptions D.1, We also assume that m is sufficiently large such that $m = \Omega(T^{13}H^{14} \cdot (\log m)^3)$. In Algorithm 4, we let λ be a sufficiently large constant and let $\beta = B_T$ which satisfies inequality

$$16\Gamma_{K_m}(T, \lambda) + 16 \cdot \log N_\infty(\epsilon^*, h+1, B_T) + 32 \cdot \log(2TH) + 4R_Q^2 \cdot (1 + \lambda/d) \le (B_T/H)^2 \tag{D.3}$$

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 (B.3), 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

$$m = \Omega\big(\beta^{12} \cdot T^{13} \cdot H^{14} \cdot (\log m)^3\big). \tag{D.4}$$

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

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

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 (D.3) 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 (D.6) and (D.5) 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 $\texttt{err}_{\mathrm{mis}} = 10\beta \cdot \iota$. Specifically, the first term in (D.5) is the same as that in (D.6), 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 $\texttt{err}_{\mathrm{mis}} \cdot TH$, 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 $\texttt{err}_{\mathrm{mis}} \cdot TH$ 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 (D.3), we simplify the regret bound in (D.5) into

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

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 D.3 (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 $\{\sigma_j\}_{j \ge 1}$ and $\{\psi_j\}_{j \ge 1}$ be its eigenvalues and eigenvectors, respectively. We assume that $\{\sigma_j\}_{j \ge 1}$ and $\{\psi_j\}_{j \ge 1}$ satisfy either one of the two 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 §G.3 in the appendix.

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

Corollary D.4. Under Assumptions D.1 and D.3, we assume the activation function is C_{act}-smooth and the number of neurons of the neural network satisfies (D.4). Besides, in Algorithm 4 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

$$\mathrm{Regret}(T) \le \begin{cases} C_r \cdot H^2 \cdot \sqrt{\gamma^3 T} \cdot \log(\gamma TH) + 10\beta TH \cdot \iota & \gamma\text{-finite spectrum}, \\ C_r \cdot H^2 \cdot \sqrt{(\log T)^{3/\gamma} \cdot T} \cdot \log(TH) + 10\beta TH \cdot \iota & \gamma\text{-exponential decay}, \end{cases} \tag{D.6}$$

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

Corollary D.4 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 §G.3, neural network with quadratic and sine activation functions induce NTKs satisfying the finite-spectrum and exponential eigenvalue decay conditions, respectively. Corollary D.4 can be directly applied to these concrete examples to obtain sublinear regret bounds.

E Proofs of the Main Results

In this section, we provide the proofs of Theorems 4.2 and D.2. The proofs of the supporting lemmas and auxiliary results are deferred to the appendix.

E.1 Proof of Theorem 4.2

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

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

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 the $\{Q_h^t\}_{h\in[H]}$ are from satisfying the Bellman optimality equation in (2.2). Next, 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

$$\zeta_{t,h}^1 = \big[V_h^t(x_h^t) - V_h^{\pi^t}(x_h^t)\big] - \big[Q_h^t(x_h^t, a_h^t) - Q_h^{\pi^t}(x_h^t, a_h^t)\big], \tag{E.2}$$

$$\zeta_{t,h}^2 = \big[(\mathbb{P}_h V_{h+1}^t)(x_h^t, a_h^t) - (\mathbb{P}_h V_{h+1}^{\pi^t})(x_h^t, a_h^t)\big] - \big[V_{h+1}^t(x_{h+1}^t) - V_{h+1}^{\pi^t}(x_{h+1}^t)\big]. \tag{E.3}$$

By definition, $\zeta_{t,h}^1$ and $\zeta_{t,h}^2$ capture two sources of randomness—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 $\mathbb{P}_h(\cdot \,|\, x_h^t, a_h^t)$, respectively. As we will see in Appendix §H.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 [7].

To establish an upper bound on the regret, the following lemma first decomposes the regret into three parts using the notation defined above. Similar regret decomposition results also appear in [12, 29].

Lemma E.1 (Regret Decomposition)**.** The temporal-difference error is the mapping $\delta_h^t : \mathcal{S} \times \mathcal{A} \to$ defined in (E.1) for all $(t,h) \in [T] \times [H]$. We can thus write the regret as

$$\mathrm{Regret}(T) = \underbrace{\sum_{t=1}^{T} \sum_{h=1}^{H} \big[\mathbb{E}_{\pi^\star}[\delta_h^t(x_h, a_h) \,|\, x_1 = x_1^t] - \delta_h^t(x_h^t, a_h^t)\big]}_{\text{(i)}} + \underbrace{\sum_{t=1}^{T} \sum_{h=1}^{H} (\zeta_{t,h}^1 + \zeta_{t,h}^2)}_{\text{(ii)}}$$
$$\underbrace{\sum_{t=1}^{T} \sum_{h=1}^{H} \mathbb{E}_{\pi^\star}\big[\big\langle Q_h^t(x_h, \cdot), \pi_h^\star(\cdot \,|\, x_h) - \pi_h^t(\cdot \,|\, x_h)\big\rangle_{\mathcal{A}} \,\big|\, x_1 = x_1^t\big]}_{\text{(iii)}}, \tag{E.4}$$

where $\zeta_{t,h}^1$ and $\zeta_{t,h}^2$ are defined in (E.2) and (E.3), respectively.

Proof. See Appendix §H.1 for a detailed proof. □

Returning to the main proof, notice that π_h^t is the greedy policy with respect to Q_h^t for all $(t,h) \in [T] \times [H]$. We have

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

for all $x_h \in \mathcal{S}$. Thus, Term (iii) in (E.4) is non-positive. Then, by Lemma E.1, we can upper bound the regret by

$$\text{Regret}(T) \leq \underbrace{\left\{ \sum_{t=1}^{T} \sum_{h=1}^{H} \left[\mathbb{E}_{\pi^\star}[\delta_h^t(x_h, a_h) \,|\, x_1 = x_1^t] - \delta_h^t(x_h^t, a_h^t) \right] \right\}}_{\text{(i)}} + \underbrace{\left[\sum_{t=1}^{T} \sum_{h=1}^{H} (\zeta_{t,h}^1 + \zeta_{t,h}^2) \right]}_{\text{(ii)}}. \tag{E.5}$$

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, however, by adding the bonus term in Algorithm 2, we ensure that the temporal-difference error δ_h^t is a non-positive function, as shown in the following lemma.

Lemma E.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 $1 - (2T^2H^2)^{-1}$, we have that the following holds for all $(t,h) \in [T] \times [H]$ and $(x,a) \in \mathcal{S} \times \mathcal{A}$:

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

Proof. See Appendix §H.2 for a detailed proof. □

Applying Lemma E.2 to Term (i) in (E.5), we obtain that

$$\text{Term (i)} \leq \left[\sum_{t=1}^{T} \sum_{h=1}^{H} -\delta_h^t(x_h^t, a_h^t) \right] \leq 2\beta \cdot \left[\sum_{t=1}^{T} \sum_{h=1}^{H} b_h^t(x_h^t, a_h^t) \right] \tag{E.6}$$

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

Finally, it remains to bound the sum of bonus terms in (E.6). As we show in (H.17), using the feature representation of $\mathcal{H}$, we can write each $b_h^t(x_h^t, a_h^t)$ as

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

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 the Cauchy-Schwarz inequality and Lemma J.3, we have, for any $h \in [H]$, with probability at least $1 - (2T^2H^2)^{-1}$ the following:

$$\begin{aligned} \text{Term (i)} &\leq 2\beta \cdot \sqrt{T} \cdot \sum_{h=1}^{H} \left[\sum_{t=1}^{T} \phi(x_h^t, a_h^t)^\top (\Lambda_h^t)^{-1} \phi(x_h^t, a_h^t) \right]^{1/2} \\ &\leq 2\beta \cdot \sum_{h=1}^{H} \left[2T \cdot \text{logdet}(I + K_h^T/\lambda) \right]^{1/2} = 4\beta H \cdot \sqrt{T \cdot \Gamma_K(T, \lambda)}, \end{aligned} \tag{E.7}$$

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

It remains to bound Term (ii) in (E.5), which is the purpose of the following lemma.

Lemma E.3. For $\zeta_{t,h}^1$ and $\zeta_{t,h}^2$ defined respectively in (E.2) and (E.3) and for any $\zeta \in (0,1)$, with probability at least $1 - \zeta$, we have

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

Proof. See Appendix §H.3 for a detailed proof. □

Setting $\zeta = (2T^2H^2)^{-1}$ in Lemma E.3 we obtain that

$$\text{Term (ii)} = \sum_{t=1}^{T}\sum_{h=1}^{H}(\zeta_{t,h}^1 + \zeta_{t,h}^2) \le \sqrt{16TH^3 \cdot \log(4T^2H^2)} = \sqrt{32TH^3 \cdot \log(2TH)} \tag{E.8}$$

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

Therefore, combining (4.5), (E.5), and (E.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

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

This concludes the proof of Theorem 4.2. □

E.2 Proof of Theorem D.2

Proof. The proof of Theorem D.2 is similar to that of Theorem 4.2. Recall that we let $\mathcal{Z}$ denote $\mathcal{S} \times \mathcal{A}$ for simplicity. Recall also that for all $(t,h) \in [T] \times [H]$, we define the temporal-difference (TD) error $\delta_h^t \colon \mathcal{Z} \to \mathbb{R}$ in (E.1) and define random variables $\zeta_{t,h}^1$ and $\zeta_{t,h}^2$ in (E.2) and (E.3), respectively.

Then, combining Lemma E.1 and the fact that π^t is the greedy policy with respect to $\{Q_h^t\}_{h\in[H]}$, we bound the regret by

$$\text{Regret}(T) \le \underbrace{\biggl\{\sum_{t=1}^{T}\sum_{h=1}^{H}\bigl[\mathbb{E}_{\pi^\star}[\delta_h^t(x_h,a_h) \,|\, x_1 = x_1^t] - \delta_h^t(x_h^t,a_h^t)\bigr]\biggr\}}_{\text{(i)}} + \underbrace{\biggl[\sum_{t=1}^{T}\sum_{h=1}^{H}(\zeta_{t,h}^1 + \zeta_{t,h}^2)\biggr]}_{\text{(ii)}}. \tag{E.9}$$

Here, Term (ii) is a sum of a martingale difference sequence. By setting $\zeta = (4T^2H^2)^{-1}$ in Lemma E.3, with probability at least $1 - (4T^2H^2)^{-1}$, we have

$$\text{Term (ii)} = \sum_{t=1}^{T}\sum_{h=1}^{H}(\zeta_{t,h}^1 + \zeta_{t,h}^2) \le \sqrt{16TH^3 \cdot \log(8T^2H^2)} \le H \cdot \sqrt{32TH\log(2TH)}. \tag{E.10}$$

It remains to bound Term (i) in (E.9). To this end, we aim to establish a counterpart of Lemma E.2 for neural value functions, which shows that, by adding a bonus term $\beta \cdot b_h^t$, the TD error δ_h^t is always a non-positive function approximately. This implies that bounding Term (i) in (E.9) reduces to controlling $\sum_{t=1}^{T}\sum_{h=1}^{H} b_h^t(x_h^t,a_h^t)$.

Note that the bonus functions b_h^t are constructed based on the neural tangent features $\varphi(\cdot;\widehat{W}_h^t)$ and the 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, we define $\overline{\Lambda}_h^t$ and $\overline{b}_h^t$, by analogy with Λ_h^t and b_h^t, as follows:

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

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 E.4 (Optimism)**.** Let λ be an absolute constant and let $\beta = B_T$ in Algorithm 4, where B_T satisfies (D.3). Under the assumptions made in Theorem D.2, 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

$$-5\beta \cdot \iota - 2\beta \cdot \overline{b}_h^t(x,a) \le \delta_h^t(x,a) \le 5\beta \cdot \iota, \qquad \sup_{(x,a)\in\mathcal{Z}}\bigl|b_h^t(x,a) - \overline{b}_h^t(x,a)\bigr| \le 2\iota, \tag{E.11}$$

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

Proof. See Appendix §H.4 for a detailed proof. □

Applying Lemma E.2 to Term (i) in (E.5), we obtain that

$$\text{Term (i)} \le \biggl[\sum_{t=1}^{T}\sum_{h=1}^{H} -\delta_h^t(x_h^t, a_h^t)\biggr] + 5TH\cdot\iota \le 2\beta\cdot\biggl[\sum_{t=1}^{T}\sum_{h=1}^{H}\overline{b}_h^t(x_h^t, a_h^t)\biggr] + 10\beta TH\cdot\iota \quad \text{(E.12)}$$

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

$$\begin{aligned}\sum_{t=1}^{T}\sum_{h=1}^{H}\overline{b}_h^t(x_h^t, a_h^t) &\le \sqrt{T}\cdot\sum_{h=1}^{H}\biggl[\sum_{t=1}^{T}\varphi(x_h^t, a_h^t; W^{(0)})^\top(\overline{\Lambda}_h^t)^{-1}\varphi(x_h^t, a_h^t; W^{(0)})\biggr]^{1/2}\\ &\le 2H\cdot\sqrt{T\cdot\Gamma_{K_m}(T,\lambda)}, \qquad \text{(E.13)}\end{aligned}$$

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 (E.9), (E.10), (E.12), and (E.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

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

This concludes the proof of Theorem D.2. □

F Neural Optimistic Least-Squares Value Iteration

In this section, we provide the pseudocode for NOVI, which was omitted in the main text for brevity. We remark that the loss function L_h^t in Line 7 is given in (C.1) and its global minimizer $\widehat{W}_h^t$ can be efficiently obtained by first-order optimization methods.

Algorithm 4 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 $\widehat{W}_h^t = \operatorname{argmin}_W L_h^t(W)$.
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 (C.4).
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 $a_h^t \leftarrow \operatorname{argmax}_{a\in\mathcal{A}} Q_h^t(x_h^t, a)$.
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**

G Proofs of the Corollaries

In this section, we prove Corollaries 4.4 and D.4, which establish the regret for KOVI and NOVI under each specific eigenvalue decay condition. in Appendix §G.3 we provide concrete examples of neural

tangent kernels that satisfy Assumption 4.3 and show how to apply Corollaries 4.4 and D.4 to these examples.

G.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 two cases separately.

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

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

where $C_N > 0$ is the absolute constant given in Lemma I.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 (G.1), the left-hand side (LHS) of (4.5) is bounded by

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

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 (G.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

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

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 I.5 we have

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

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 I.1 we have

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

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

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

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

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

where $\overline{C}_3$ is an absolute constant that does not depend on C_b. Thus, when C_b is sufficiently large, (G.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 (G.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. Thus we conclude the second case. Therefore, we conclude the proof of Corollary 4.4. □

G.2 Proof of Corollary D.4

Proof. By Theorem D.2, we have

$$\mathrm{Regret}(T) = 5\beta H \cdot \sqrt{T \cdot \Gamma_{K_m}(T, \lambda)} + 10\beta T H \cdot \iota, \tag{G.6}$$

where $\beta = B_T$ satisfies (D.3) and $\iota = T^{7/12} \cdot H^{1/6} \cdot m^{-1/12} \cdot (\log m)^{1/4}$. When Assumption D.3 holds, thanks to the similarity between (4.5) and (D.3), it can be similarly shown that B_T defined in (4.8) satisfies the inequality in (D.3) when C_b is sufficiently large. Moreover, Lemma I.5 provides upper bounds on $\Gamma_{K_m}(T, \lambda)$ for the two eigenvalue decay conditions. Finally, combining (4.8), (G.6), and Lemma I.5, we conclude the proof of Corollary D.4. □

G.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 $\{z \in \mathbb{R}^d \colon \|z\|_2 \le 1\}$, where n is a fixed number. By direct computation, the kernel function can be written as

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

where $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$ 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 satisfies 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 network setting. Moreover, in (B.3), 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 $|W_j^\top z| \le 1$ 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 $\{W_j^\top z \colon z \in \mathbb{S}^{d-1}\} \subseteq [-1, 1]$ for any W_j sampled from the initial distribution, which is required when utilizing Proposition C.1 in [30] in the proof of Lemma E.4. Note that the covariance of W_j is still I_d/d. Then by (B.7), the NTK is given by

$$K_{\mathrm{ntk}}(z, z') = \mathbb{E}_{w \sim \mathrm{Unif}(\mathbb{S}^{d-1})}[2(w^\top z) \cdot 2(w^\top z') \cdot (z^\top z')] = 4/d \cdot (z^\top z')^2, \qquad \forall z, z' \in \mathbb{S}^{d-1}. \tag{G.7}$$

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 $j \ge 0$, 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

$$N(d, j) = \frac{(2j + d - 2)(d + j - 3)!}{j!(d - 2)!}. \tag{G.8}$$

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 [48, page 30], which relates spherical harmonics to inner product kernels.

Lemma G.1 (Funk-Hecke formula)**.** Let $k\colon [-1,1] \to \mathbb{R}$ be a continuous function, which gives 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. Moreover, for any $j \geq 0$, let $Y_j\colon \mathbb{S}^{d-1} \to \mathbb{R}$ be any function in $\mathcal{Y}_j(d)$. Then for any $z \in \mathbb{S}^{d-1}$, we have

$$\int_{\mathbb{S}^{d-1}} K(z,z') Y_j(z') \,\mathrm{d}\mu(z') = \bigg[\frac{|\mathbb{S}^{d-2}|}{|\mathbb{S}^{d-1}|} \cdot \int_{-1}^{1} k(u) \cdot P_j(u;d) \cdot (1-u^2)^{(d-3)/2} \,\mathrm{d}u\bigg] \cdot Y_j(z), \tag{G.9}$$

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

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

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

$$\varrho_j = \frac{|\mathbb{S}^{d-2}|}{|\mathbb{S}^{d-1}|} \cdot \int_{-1}^{1} k(u) \cdot P_j(u;d) \cdot (1-u^2)^{(d-3)/2} \,\mathrm{d}u, \qquad \forall j \geq 1, \tag{G.10}$$

each with multiplicity $N(d,j)$. Moreover, 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 (G.10), we can use Rodrigues' rule [48, page 23], as follows.

Lemma G.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 (G.7), which is the inner product kernel induced by the univariate function $k_1(u) = 4/d \cdot u^2$. Note that $k_1^{(3)}$ is a zero function. Combining Lemma G.2 and (G.10), we observe that $\varrho_j = 0$ for all $j \geq 3$. In addition, by direct computation, we have 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 $\varrho_0, \varrho_2 > 0$. Thus, K_{ntk} given in (G.7) 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 the γ-finite spectrum condition with $\gamma = d(d+1)/2$. For such a class of neural networks, Corollary D.4 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

$$K(z,z') = \exp(-\|z-z'\|_2^2 \cdot \sigma^{-2}) = k_2(\langle z, z'\rangle), \qquad \forall z, z' \in \mathbb{S}^{d-1}, \tag{G.11}$$

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 G.1 and G.2, we obtain the following lemma that bounds the eigenvalues of T_K.

Lemma G.3 (Theorem 2 in [47])**.** For the squared quadratic kernel in (G.11), the corresponding integral operator has eigenvalues $\{\rho_j\}_{j \geq 0}$, where each ρ_j is defined in (G.10) with k replaced by k_2. Moreover, each ϱ_j has multiplicity $N(d,j)$ and the corresponding eigenfunctions are $\{Y_{j,\ell}\}_{\ell \in [N(d,j)]}$. Finally, when σ in (G.11) satisfy $\sigma^2 \geq 2/d$, $\{\varrho_j\}_{j \geq 0}$ form a decreasing sequence that satisfy

$$A_1 \cdot (2e/\sigma^2)^j \cdot (2j+d-2)^{-(2j+d-1)/2} < \varrho_j < A_2 \cdot (2e/\sigma^2)^j \cdot (2j+d-2)^{-(2j+d-1)/2} \tag{G.12}$$

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

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

Lemma G.4 (Lemma 3 in [47]). 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 G.3 and G.4, we have

$$\varrho_j^\tau \cdot \|Y_{j,\ell}\|_\infty \leq C \cdot \Big(\frac{2e}{\sigma^2 \cdot (2j+d-2)}\Big)^{-j\cdot\tau} \cdot \sqrt{N(d,j) \cdot (2j+d-2)^{-(d-1)\cdot\tau}}, \tag{G.13}$$

where C is a constant depending on d and σ. By the definition of $N(d,j)$ in (G.8), when j is sufficiently large, it holds that

$$N(d,j) \asymp \frac{(2j+d-2)\cdot\sqrt{d+j-3}\cdot[(d+j-3)/e]^{d+j-3}}{\sqrt{j}\cdot(j/e)^j} \asymp j^{d-2}, \tag{G.14}$$

where we utilize the Stirling's formula and neglect constants involving d. Then, combining (G.13) and (G.14), we have

$$\sup_{j\geq 0} \sup_{\ell \in [N(d,j)]} \varrho_j^\tau \cdot \|Y_{j,\ell}\|_\infty \leq C_\varrho, \tag{G.15}$$

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, (G.15) 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

$$\sigma_j = \varrho_t \qquad \text{for} \quad \sum_{i=1}^{t-1} N(d,i) \leq j < \sum_{i=1}^{t} N(d,i). \tag{G.16}$$

Then by (G.14), 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 G.3 we further obtain that

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

where c, c_1, and c_2 are constants depending on d. Therefore, we have shown that the squared exponential kernel satisfies the γ-exponential decay condition with $\gamma = 1/d$. Combining this with (G.15), we conclude that it satisfies Assumption 4.3.

In the sequel, we construct an NTK that satisfies Assumption 4.3. Specifically, we adopt the sine activation function and slightly modify the neural network in (B.3) by employing an intercept for 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 $b_j \sim \text{Unif}(\{-1, 1\})$, $W_j \sim N(0, I_d)$, and $\theta_j \sim \text{Unif}([0, 2\pi])$. Only W is updated during training.

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

$$\begin{aligned}
K_{\text{ntk}}(z, z') &= 2\mathbb{E}\big[(z^\top z') \cdot \cos(w^\top z + \theta) \cdot \cos(w^\top z' + \theta)\big] \\
&= (z^\top z') \cdot \exp(-\|z - z'\|_2^2/2) = (z^\top z') \cdot \exp[(z^\top z') - 1] = k_3(\langle z, z'\rangle),
\end{aligned} \tag{G.17}$$

where we define $k_3(u) = u \cdot \exp(u-1)$. Here the second equality follows from [54]. By construction, such an NTK is closely related to the squared quadratic kernel in (G.11). 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 (G.17) and the inner product kernel induced by $k_2(u) = \exp(u-1)$, respectively. By Lemma G.1, we have

$$\begin{aligned}
\rho_j &= C_1 \cdot \int_{-1}^{1} k_3(u) \cdot P_j(u; d) \cdot (1-u^2)^{(d-3)/2} \, \mathrm{d}u = C_1 \cdot \int_{-1}^{1} k_2(u) \cdot u \cdot P_j(u; d) \cdot (1-u^2)^{(d-3)/2} \, \mathrm{d}u \\
&= C_2 \cdot j/(2j+d-2) \cdot \widetilde{\varrho}_{j-1} + C_2 \cdot (j+d-2)/(2j+d-2) \cdot \widetilde{\varrho}_{j+1} \leq C_2(\widetilde{\rho}_{j-1} + \widetilde{\rho}_{j+1}),
\end{aligned} \tag{G.18}$$

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 (G.12). Thus, combining (G.12) and (G.18), we obtain (G.15). Moreover, when ordering all the eigenvalues of K_{ntk} in the descending order and renaming them as $\{\sigma_j\}_{j\geq 1}$, similar to (G.16), we have

$$\sigma_j \leq C_2 \cdot (\widetilde{\rho}_{t-1} + \widetilde{\rho}_{t+1}) \qquad \text{for } \sum_{i=1}^{t-1} N(d,i) \leq j < \sum_{i=1}^{t} N(d,i). \tag{G.19}$$

Using a 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 (G.17) satisfy Assumption 4.3.

H Proofs of the Supporting Lemmas

H.1 Proof of Lemma E.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

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

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

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

After introducing this notation, to prove (E.4) we decompose the instantaneous regret at the t-th episode into the following two terms,

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

In the sequel, we consider the two terms in (H.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 (H.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 (H.1), we have $V_h^t = \mathbb{J}_{t,h} Q_h^t$. Thus, using ξ_h^t defined in (H.2), for any $(t,h) \in [T] \times [H]$, we have

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

where the last equality follows from the definition of ξ_h^t in (H.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 (E.1) and the Bellman optimality condition, we have

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

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

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

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 (H.6) for all $h \in [H]$, we have

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

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

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

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

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

Term (ii). It remains to bound the second term on the right-hand side of (H.3). By the definition of the temporal-difference error δ_h^t in (E.1), for any $(t,h)\in[T]\times[H]$, we have

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

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 (E.2) and (E.3), respectively. Thus, from (H.9) we obtain that

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

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

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

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 (H.3).

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

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

Therefore, we conclude the proof of this lemma. □

H.2 Proof of Lemma E.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

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

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

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

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 (H.13) and y_h^t defined in (3.6), we can write (H.12) as

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

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

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

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

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

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

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

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

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

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

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

Now we are ready to bound the temporal-difference error ξ_h^t defined in (E.1). Noticing 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 (H.16), we can write $\phi(z)^\top \overline{\theta}_h^t$ as

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

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

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

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

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

Here the first inequality follows from the Cauchy-Schwarz inequality and the second inequality follows from the fact that $\overline{\theta}_h^t \in \mathcal{Q}^\star$, which implies that $\|\overline{\theta}_h^t\|_{\mathcal{H}} \le R_Q H$. Moreover, 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 (H.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 $\big(y_h^t - \Phi \overline{\theta}_h^t\big)$ can be written as

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

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

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

where the last inequality follows from the Cauchy-Schwarz inequality. In the following, we aim to bound (H.23) by the concentration of self-normalized stochastic processes 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 processes and uniform convergence over the function classes that contain each V_{h+1}^t.

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

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

In the following, we find a parameter R_T such that $V_h^t \in \mathcal{V}_{\mathrm{ucb}}(h, R_T, B_T)$ 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 (H.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, we bound the RKHS norm of each $\widehat{\theta}_h^t$.

Lemma H.1 (RKHS Norm of $\widehat{\theta}_h^t$)**.** When $\lambda \geq 1$, for any $(t,h) \in [T] \times [H]$, $\widehat{\theta}_h^t$ defined in (H.15) satisfies

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

where K_h^t is defined in (3.7) and $\Gamma_K(T,\lambda)$ is defined in (I.16).

Proof. See §J.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 (H.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 $\mathrm{dist}(V, V') = \sup_{x \in \mathcal{S}} |V(x) - V'(x)|$. For any $\epsilon \in (0, 1/e)$, any $B > 0$, and any $h \in [H]$, we let $N_{\mathrm{dist}}(\epsilon; h, B)$ be the ϵ-covering number of $\mathcal{V}_{\mathrm{ucb}}(h, R_T, B)$ with respect to distance $\mathrm{dist}(\cdot,\cdot)$. Recall that we define $N_\infty(\epsilon; h, B)$ as the ϵ-covering number of $\mathcal{Q}_{\mathrm{ucb}}(h, R_T, B)$ with respect to the ℓ_∞-norm on $\mathcal{Z}$. Note that for any $Q, Q' \colon \mathcal{Z} \to \mathbb{R}$, we have

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

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

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

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 $V_{h+1}^t \in \mathcal{V}_{\mathrm{ucb}}(h+1, R_T, B_T)$. Note that we set $\lambda = 1 + 1/T$. Then, setting ϵ as $\epsilon^* = H/T$, (H.25) is further reduced to

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

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

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

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 elementary inequality $\sqrt{a} + \sqrt{b} \leq \sqrt{2(a^2 + b^2)}$, and the last inequality follows from the assumption on B_T given in (4.5).

Finally, by (H.27) and the definition of the temporal-difference error δ_h^t in (E.1), we have

$$-\delta_h^t(z) = Q_h^t(z) - \phi(z)^\top \overline{\theta}_h^t \leq \phi(z)^\top (\widehat{\theta}_h^t - \overline{\theta}_h^t) + \beta \cdot b_h^t(z) \leq 2\beta \cdot b_h^t(z). \qquad \text{(H.28)}$$

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

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

Therefore, combining (H.28) and (H.29), we conclude the proof of Lemma E.2. □

H.3 Proof of Lemma E.3

Proof. Following [12], 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 $\mathcal{F}_{t,h,2}$ as follows:

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

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

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

which offers an partial ordering over the triplets $(t, h, m) \in [T] \times [H] \times [2]$. Moreover, by the definitions in (H.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 of 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 (H.30), $\zeta_{t,h}^1$ defined in (E.2) and $\zeta_{t,h}^2$ defined in (E.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

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

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 (H.31) and (H.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 (H.31), as follows. For any $(t, h, m) \in [T] \times [H] \times [2]$, we define

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

that is, $M_{t,h,m}$ is the sum of all terms of the form $\zeta_{s,g}^\ell$ defined in (E.2) or (E.3) such that its timestep index $\tau(s, g, \ell)$ is no greater than $\tau(t, h, m)$. By definition, we have

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

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 (H.33) has uniformly bounded differences. Thus, applying the Azuma-Hoeffding inequality [7] to $M_{T,H,2}$ in (H.34), we obtain that

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

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

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

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

H.4 Proof of Lemma E.4

Proof. The proof of this lemma utilizes the connection between overparameterized neural networks and NTKs. Recall 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 §B.2. 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 (C.1), we have

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

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, (H.36) implies that

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

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 D.1, 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 (D.1). 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

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

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 (E.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 [30], 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

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

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

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

Also, 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 (H.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

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

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

Therefore, for all $(t,h)\in[T]\times[H]$, we have constructed $\widetilde{Q}_h^t$ to be linear in $\varphi(\cdot;W^{(0)})$. 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 (H.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 that Q_h^t used in Algorithm 4 can be well approximated by 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 (C.3), 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

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

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. Moreover, 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

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

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

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

Hence, by (H.42) and (H.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 (C.4), we define

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

Similar to L_h^t defined in (C.1), we define another least-squares loss function $\overline{L}_h^t\colon\mathbb{R}^{2md}\to\mathbb{R}$ as

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

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

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

where $\overline{\Lambda}_h^t$, Φ_h^t, and y_h^t are defined respectively in (H.42), (H.44), and (H.43). Similar to (H.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, in a manner similar to the construction of Q_h^t in Algorithm 4, we combine $\overline{b}_h^t$ in (H.45) and $\overline{W}_h^t$ in (H.47) to define $\overline{Q}_h^t\colon\mathcal{Z}\to\mathbb{R}$ as

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

Note that $\overline{Q}_h^t$ share the same form as Q in (D.2). Thus, we have $\overline{Q}_h^t \in \mathcal{Q}_{\mathrm{ucb}}(h, H\sqrt{T/\lambda}, B)$ for any $B \geq \beta$. 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 the triangle inequality we have

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

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

Lemma H.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

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

Proof. See [3, 30, 13] for a detailed proof. More specifically, this lemma is obtained from Lemmas F.1 and F.2 in [13], which are further based on results in [3, 30]. □

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

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

To bound $\big\|\widehat{W}_h^t - \overline{W}_h^t\big\|_2$, notice that $\widehat{W}_h^t$ and $\overline{W}_h^t$ are the global minimizers of L_h^t in (C.1) and $\overline{L}_h^t$ in (H.46), respectively. Thus, by the first-order optimality condition, we have

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

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

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

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

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

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

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

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

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

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

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

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

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 (H.57) dominates. Since the eigenvalues of $(\overline{\Lambda}_h^t)^{-1}$ are all bounded by $1/\lambda$, we have

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

In the sequel, we set λ as

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

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

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

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

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

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

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

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

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

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

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

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

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

By direct computation, we have

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

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

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

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

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

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

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

which establishes the second inequality in (E.11). Finally, combining (H.49), (H.60), and (H.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

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

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 the function $\widetilde{Q}_h^t$ defined in (H.39), where $\varphi(\cdot)$ denotes $\varphi(\cdot;W^{(0)})$. By the definition of $\overline{\Lambda}_h^t$ in (H.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 (H.41). Hence, combining (H.47), we have

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

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

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

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

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

For Term (iv) in (H.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 (H.40), for any $\tau\in[t-1]$, the τ-th entry of $\Phi_h^t(\widetilde{W}_h^t-W^{(0)})$ satisfies

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

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

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

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

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

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

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

Moreover, we define functions $\Delta V_1=V_{h+1}^t-\overline{V}_{h+1}^t$ and $\Delta V_2=\mathbb{P}_h(V_{h+1}^t-\overline{V}_{h+1}^t)$. Then (H.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 elementary inequality $(a+b)^2\le 2a^2+2b^2$, we have

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

where the last inequality follows from the fact that

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

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

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

$$\begin{aligned}
&\big|\varphi(z)^\top(\overline{\Lambda}_h^t)^{-1}(\Phi_h^t)^\top\big[y_h^t - \Phi_h^t\big(\widetilde{W}_h^t - W^{(0)}\big)\big]\big| \\
&\quad\le \overline{b}_h^t(z)\cdot\big[10C_{\mathrm{act}}\cdot R_Q H\cdot\sqrt{\log(mTH)/m} + \sqrt{2\cdot\text{Term (v)}} + 2\sqrt{2}\cdot\texttt{Err}\cdot T\big] \\
&\quad\le \overline{b}_h^t(z)\cdot\big[R_Q H + \sqrt{2\cdot\text{Term (v)}}\big], \qquad \text{(H.77)}
\end{aligned}$$

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

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

In the following, we aim to bound Term (v) in (H.77) by combining the concentration of the self-normalized stochastic process and uniform concentration. To characterize the function class that contains each $\overline{V}_h^t$, 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 H.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

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

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 (H.59), we have $\widetilde{\lambda} = 1 + 1/T\in[1,2]$. Moreover, by (H.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 (H.47), we have

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

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

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

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

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

For any $h\in[H]$, $R,B>0$, we let $\widetilde{\mathcal{Q}}_{\mathrm{ucb}}(h,R,B)$ 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 $\widetilde{\mathcal{Q}}_{\mathrm{ucb}}(h,R,B)$ corresponds to the function class in (4.4) with the kernel being $\widetilde{K}$. Moreover, 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 $R,B>0$, we have 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 $R>0$, we have

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

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 (D.3). Since we set $\beta = B_T$ in Algorithm 4, 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 (H.77), similar to the analysis the proof of Lemma E.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 (H.25) and (H.26), with probability at least $1 - (2T^2H^2)^{-1}$, we have

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

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

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

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

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

By the choice of B_T in (D.3), we have that

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

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 (H.39) and (H.40), we have

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

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

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

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

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

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

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

Meanwhile, combining the definition of the TD error δ_h^t in (E.1) and (H.69), we have

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

Finally, combining (H.85), (H.86), and (H.87), we establish that, with probability at least $1 - (2T^2H^2)^{-1}$,

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

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

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

which, together with (H.68), concludes the proof of Lemma E.4. □

I Covering Number and Effective Dimension

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

I.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 following form:

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

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,

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

By definition, Q in (I.1) is determined by $Q_0 \in \mathcal{H}$ and a bonus term constructed using $\mathcal{D}$. Thus, the 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 the function class $\mathcal{V}_{\text{ucb}}(h, R, B)$ as

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

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

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

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 (I.3).

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

Lemma I.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

$$\log\big|\mathcal{C}\big(\mathcal{V}_{\text{ucb}}(h, R, B), \epsilon\big)\big| \le \log\big|\mathcal{C}\big(\mathcal{Q}_{\text{ucb}}(h, R, B), \epsilon\big)\big| \quad \text{(I.4)}$$

$$\le \begin{cases} C_N \cdot \gamma \cdot \big[1 + \log(R/\epsilon)\big] + C_N \cdot \gamma^2 \cdot \big[1 + \log(B/\epsilon)\big] & \text{case (i)}, \\ C_N \cdot \big[1 + \log(R/\epsilon)\big]^{1+1/\gamma} + C_N \cdot \big[1 + \log(B/\epsilon)\big]^{1+2/\gamma} & \text{case (ii)}, \end{cases} \quad \text{(I.5)}$$

where cases (i) and (ii) above correspond to the two eigenvalue decay conditions specified in Assumption 4.3, respectively. Moreover, here C_N in (I.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

$$\Phi_\mathcal{D} = \big[\phi(z^1)^\top, \phi(z^2)^\top \ldots, \phi(z^t)^\top\big]^\top,$$
$$\Lambda_\mathcal{D} = \sum_{\tau=1}^{t} \phi(z^\tau)\phi(z^\tau)^\top + \lambda \cdot I_\mathcal{H} = \lambda \cdot I_\mathcal{H} + (\Phi_\mathcal{D})^\top \Phi_\mathcal{D}, \tag{I.6}$$

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 $Q_1 \in \mathcal{Q}_{\mathrm{ucb}}(h, R, B)$ as

$$Q_1(z) = \phi(z)^\top\theta_1 + \beta \cdot \sqrt{\phi(z)^\top \Lambda_{\mathcal{D}_1}^{-1}\phi(z)}, \tag{I.7}$$

where $\theta_1 \in \mathcal{H}$ has an RKHS norm 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 (I.2) and (I.7), we can write $V_1 \in \mathcal{V}_{\mathrm{ucb}}(h, R, B)$ as

$$V_1(\cdot) = \min\Big\{\max_a\big\{\phi(\cdot, a)^\top\theta_1 + \beta_1 \cdot \sqrt{\phi(\cdot, a)^\top \Lambda_{\mathcal{D}_1}^{-1}\phi(\cdot, a)}\big\}, H - h + 1\Big\}^+, \tag{I.8}$$

Similar to V_1 in (I.8), consider any $V_2\colon \mathcal{S} \to \mathbb{R}$ that can be written as

$$V_2(\cdot) = \min\Big\{\max_a\big\{f_1(\cdot, a) + \beta_2 \cdot f_2(\cdot, a)\big\}, H - h + 1\Big\}^+, \tag{I.9}$$

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 (I.8) and (I.9) we have

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

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 the triangle inequality, we have

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

where we denote $\big\|\phi(x, a)\big\|^2_{\Lambda_{\mathcal{D}_1}^{-1}} = \phi(x, a)^\top \Lambda_{\mathcal{D}_1}^{-1}\phi(x, a)$. 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 (I.10) we have

$$\begin{aligned}\|Q_1 - Q_2\|_\infty \le \sup_{(x,a)\in\mathcal{Z}} \big|\theta_1(x, a) - f_1(x, a)\big| + |\beta_1 - \beta_2|/\lambda \\ + B \cdot \sup_{(x,a)\in\mathcal{Z}} \big|\big\|\phi(x, a)\big\|_{\Lambda_{\mathcal{D}_1}^{-1}} - f_2(x, a)\big|.\end{aligned} \tag{I.11}$$

Thus, by (I.11), to get the covering number of $\mathcal{Q}_{\mathrm{ucb}}(h, R, B)$ with respect to $\mathrm{dist}(\cdot, \cdot)$, it suffices to bound the covering numbers of the RKHS norm ball $\{f \in \mathcal{H}\colon \|f\|_\mathcal{H} \le R\}$, the 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 (I.6), $\Lambda_\mathcal{D}\colon \mathcal{H} \to \mathcal{H}$ is a self-adjoint operator on $\mathcal{H}$ with eigenvalues bounded in $[0, 1/\lambda]$. To simplify the notation, we define the function class $\mathcal{F}(\lambda)$ as

$$\mathcal{F}(\lambda) = \big\{\|\phi(\cdot)\|_\Upsilon = \big[\phi(\cdot)^\top\Upsilon\phi(\cdot)\big]^{1/2}\colon \|\Upsilon\|_{\mathrm{op}} \le 1/\lambda\big\}, \tag{I.12}$$

where $\Upsilon\colon \mathcal{H} \to \mathcal{H}$ in (I.12) 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 (I.6). For any $\epsilon \in (0, 1)$, we let $N_\infty(\epsilon, \mathcal{F}, \lambda)$ denote the ϵ-covering number of $\mathcal{F}(\lambda)$ in (I.12)

with respect to the ℓ_∞-norm. Moreover, 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 (I.11) we obtain that

$$\left|\mathcal{C}\big(\mathcal{Q}_{\mathrm{ucb}}(h, R, B), \epsilon\big)\right| \le N_\infty(\epsilon/3, \mathcal{H}, R) \cdot N(\epsilon \cdot \lambda/3, B) \cdot N_\infty\big(\epsilon/(3B), \mathcal{F}, \lambda\big). \tag{I.13}$$

As shown in [69, Corollary 4.2.13], it holds that

$$N(\epsilon \cdot \lambda/3, B) \le 1 + 6B/(\epsilon \cdot \lambda) \le 1 + 6B/\epsilon, \tag{I.14}$$

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 (I.13) separately. We establish the ℓ_∞-covering of the RKHS norm ball and $F(\lambda)$ in the following two lemmas, respectively.

Lemma I.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 two eigenvalue decay conditions given in Assumption 4.3. Then, under Assumption 4.3, there exist absolute constants C_3 and C_4 such that

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

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 §J.2 for a detailed proof. □

Lemma I.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 (I.12). 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

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

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 §J.3 for a detailed proof. □

Finally, we conclude the proof by combining Lemmas I.2 and I.3. Specifically, by (I.13) and (I.14), we have

$$\begin{aligned} \log\left|\mathcal{C}\big(\mathcal{Q}_{\mathrm{ucb}}(h, R, B), \epsilon\big)\right| &\le \log N_\infty(\epsilon/3, \mathcal{H}, R) + \log N(\epsilon \cdot \lambda/3, B) + \log N_\infty\big(\epsilon/(3B), \mathcal{F}, \lambda\big) \\ &\le \log\big[1 + 6B/(\epsilon \cdot \lambda)\big] + \log N_\infty(\epsilon/3, \mathcal{H}, R) + \log N_\infty\big(\epsilon/(3B), \mathcal{F}, \lambda\big). \end{aligned} \tag{I.15}$$

We consider the two eigenvalue decay conditions separately. For the γ-finite spectrum case, by Lemmas I.2 and I.3 and (I.15) we have

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

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

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

for some absolute constant $C_N > 0$. Therefore, we conclude the proof. □

I.2 Effective Dimension of RKHS

Definition I.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

$$\Gamma_K(T, \sigma^2) = \sup_{\mathcal{D} \subseteq \mathcal{Z}} \big\{1/2 \cdot \operatorname{logdet}(I + \sigma^{-2} \cdot K_{\mathcal{D}})\big\}, \tag{I.16}$$

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 draw from the Gaussian process with covariance kernel K. Let $\mathcal{D} = \{z_1, \ldots, z_{|\mathcal{D}|}\}$ be a subset of $\mathcal{Z}$ with $|\mathcal{D}| \leq T$ 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 a 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

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

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 dimension of the RKHS. Specifically, as shown in [62], 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 that satisfies the exponential eigenvalue decay condition, the maximal information gain is $\mathcal{O}\big((\log T)^{d+1}\big)$. In the following lemma, similar to Theorem 5 in [62], we establish upper bounds on the maximal information gain of the RKHS under the eigenvalue decay conditions specified in Assumption 4.3.

Lemma I.5 (Theorem 5 in [62])**.** 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 (B.1). Let $\{\sigma_j\}_{j \geq 1}$ be the eigenvalues of T_K in the 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.

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

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

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 I.5 is a generalization of Theorem 5 in [62], 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 I.5 implies that the $\Gamma_K(T, \sigma^2) = \mathcal{O}((\log T)^{d+1})$, which matches Theorem 5 in [62].

Proof. The proof of this lemma is based on a modification of that of Theorem 5 in [62]. 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 [62] to bound $\Gamma_K(T, \sigma^2)$ using function B_K.

Lemma I.6 (Theorem 8 in [62])**.** Under the same condition as in Lemma I.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 [62] for a detailed proof. □

In the following, we choose proper τ and $T_\star$ in Lemma I.6 for the two 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 I.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 I.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

$$B_K(T_\star) = \sum_{j>T_\star} \sigma_j \le C_1 \cdot \sum_{j>T_\star} \exp(-C_2 \cdot j^\gamma) \le C_1 \cdot \int_{T_\star}^{\infty} \exp(-C_2 \cdot u^\gamma)\, \mathrm{d}u. \tag{I.17}$$

In a manner similar to the derivation of (J.16), by direct computation we have

$$\int_{T_\star}^{\infty} \exp(-C_2 \cdot u^\gamma)\, \mathrm{d}u \le \begin{cases} C_2^{-1} \cdot \exp(-C_2 \cdot T_\star^\gamma), & \text{if } \gamma \ge 1, \\ 2 \cdot (\gamma \cdot C_2)^{-1} \cdot \exp(-C_2 \cdot T_\star^\gamma) \cdot T_\star^{1-\gamma}, & \text{if } \gamma \in (0, 1). \end{cases} \tag{I.18}$$

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

$$\log(T \cdot n_T) = \log(C_d) + \log \cdot (T^{d+1} \cdot \log T) \le \log(C_d) + 2(d + 1) \cdot \log T, \tag{I.19}$$

when T is sufficiently large. Moreover, combining Lemma I.6 and (I.19), when σ is sandwiched by absolute constants c_1 and c_2, we have

$$\Gamma_K(T, \sigma^2) \le \widetilde{C}_1 \cdot T_\star \cdot \log T + \widetilde{C}_2 \cdot \log T \cdot \left[T^{d+1} \cdot B_K(T_\star) + 1\right] + \widetilde{C}_3, \tag{I.20}$$

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

$$\exp(C_2 \cdot T_\star^\gamma) \asymp T \cdot n_T = C_d \cdot T^{d+1} \cdot \log T, \tag{I.21}$$

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 (I.17), (I.18), and (I.21), for $\gamma \ge 1$, we have

$$\begin{aligned} &\log T \cdot \left[T^{d+1} \cdot B_K(T_\star) + 1\right] \\ &\quad \le C_1 \cdot C_2^{-1} \log T \cdot T^{d+1} \cdot \exp(-C_2 \cdot T_\star^\gamma) + \log T \le 2 \log T, \end{aligned} \tag{I.22}$$

where the last inequality follows from (I.21). Similarly, for $\gamma \in (0, 1)$, by (I.17), (I.18), and (I.21), we have

$$\begin{aligned} &\log T \cdot \left[T^{d+1} \cdot B_K(T_\star) + 1\right] \\ &\quad \le 2C_1 \cdot (\gamma \cdot C_2)^{-1} \cdot \exp(-C_2 \cdot T_\star^\gamma) \cdot \log T \cdot T^{d+1} \cdot T_\star^{1-\gamma} + \log T \asymp (\log T)^{1/\gamma - 1} + \log T. \end{aligned} \tag{I.23}$$

Thus, combining (I.20), (I.22), (I.23), we conclude that

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

for any $\gamma \ge 0$, 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. Therefore, we conclude the proof of Lemma I.5. □

J Proofs of Auxiliary Results

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

J.1 Proof of Lemma H.1

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

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

where we let Φ denote Φ_h^t defined in (H.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 (J.1) and the Cauchy-Schwarz inequality, we have

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

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 J.3, we have

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

Thus, combining (J.2), (J.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 (I.16), we conclude the proof of this lemma. □

J.2 Proof of Lemma I.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 (B.1), 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

$$f = \sum_{j=1}^{\infty} w_j \cdot \sqrt{\sigma_j} \cdot \psi_j, \tag{J.4}$$

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 (J.4). Then we have

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

In the following, we consider the two 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, (J.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 $\{w \in \mathbb{R}^\gamma \colon \|w\|_2 \le R\}$ 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 the Cauchy-Schwarz inequality, we have

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

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 $\{f \in \mathcal{H} \colon \|f\|_{\mathcal{H}} \le R\}$ is bounded by the cardinality of $\mathcal{C}_\gamma(\epsilon, R)$. As shown in [69, Corollary 4.2.13], we have

$$\big|\mathcal{C}_\gamma(\epsilon, R)\big| \le (1 + 2R/\epsilon)^\gamma. \qquad \text{(J.7)}$$

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

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

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 (J.5) we have

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

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 the Cauchy-Schwarz inequality to (J.8), we have

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

where the second inequality follows from the fact that $\sum_{j\ge1} w_j^2 \le R^2$. Since $\gamma > 0$, $\exp(-u^\gamma)$ is monotonically decreasing in u. Thus, we have

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

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

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

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

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

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

$$\frac{1}{\gamma}\int_{m^\gamma}^{\infty}\exp(-C_{2,\tau}\cdot v)\cdot v^{1/\gamma-2}\,\mathrm{d}v \le \frac{1}{m^\gamma}\cdot\frac{1}{\gamma}\int_{m^\gamma}^{\infty}\exp(-C_{2,\tau}\cdot v)\cdot v^{1/\gamma-1}\,\mathrm{d}v$$
$$= \frac{1}{m^\gamma}\int_{m}^{\infty}\exp(-C_{2,\tau}\cdot u^\gamma)\,\mathrm{d}u, \tag{J.13}$$

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

$$\int_{m}^{\infty}\exp(-C_{2,\tau}\cdot u^\gamma)\,\mathrm{d}u$$
$$\le \frac{1}{\gamma\cdot C_{2,\tau}}\cdot\exp(-C_{2,\tau}\cdot m^\gamma)\cdot m^{1-\gamma} + \frac{1/\gamma-1}{C_{2,\tau}\cdot m^\gamma}\cdot\int_{m}^{\infty}\exp(-C_{2,\tau}\cdot u^\gamma)\,\mathrm{d}u. \tag{J.14}$$

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

$$\int_{m}^{\infty}\exp(-C_{2,\tau}\cdot u^\gamma)\,\mathrm{d}u \le \left(1-\frac{1/\gamma-1}{C_{2,\tau}m^\gamma}\right)^{-1}\cdot\frac{1}{\gamma\cdot C_{2,\tau}}\exp(-C_{2,\tau}\cdot m^\gamma)\cdot m^{1-\gamma}$$
$$\le \frac{2}{\gamma\cdot C_{2,\tau}}\exp(-C_{2,\tau}\cdot m^\gamma)\cdot m^{1-\gamma}. \tag{J.15}$$

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

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

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

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

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

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

Finally, it remains to approximate $\Pi_{m^*}(f)$ up to error $\epsilon/2$ for m^* specified in (J.17). By the expansion of f in (J.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 the Cauchy-Schwarz inequality, we have

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

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 (J.19) we have

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

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 [69, Corollary

4.2.13], we have

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

Therefore, combining (J.18) and (J.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. Thus we conclude the proof of this lemma. □

J.3 Proof of Lemma I.3

Proof. As shown in §B.1, the feature mapping $\phi\colon \mathcal{Z} \to \mathcal{H}$ satisfies

$$\phi(z) = \sum_{j=1}^{\infty} \sigma_j \cdot \psi_j(z) \cdot \psi_j = \sum_{j=1}^{\infty} \sqrt{\sigma_j} \cdot \psi_j(z) \cdot (\sqrt{\sigma_j} \cdot \psi_j). \tag{J.22}$$

That is, when expanding $\phi(z) \in \mathcal{H}$ in the basis $\{\sqrt{\sigma_j} \cdot \psi_j\}_{j \ge 0}$ as in (J.4), the j-th coefficient is equal to $\sqrt{\sigma_j} \cdot \psi_j(z)$ for all $j \ge 1$. Similar to the proof of Lemma I.2, in the following, we consider the two 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 $\|\Upsilon\|_{\mathrm{op}} \le 1/\lambda$, 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

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

By (J.22) and the definition of A_Υ, we have

$$\|\phi(z)\|_\Upsilon^2 = \sum_{j,k=1}^{\gamma} \sqrt{\sigma_j} \cdot \psi_j(z) \cdot \sqrt{\sigma_k} \cdot \psi_k(z) \cdot [A_\Upsilon]_{j,k} = w_z^\top A_\Upsilon w_z. \tag{J.23}$$

With a slight abuse of notation, we define $\mathcal{C}_\gamma(\epsilon, \lambda)$ denote the minimal ϵ^2-cover of

$$\big\{A \in \mathbb{R}^{\gamma \times \gamma}\colon \|A\|_{\mathrm{fro}} \le \sqrt{\gamma}/\lambda\big\}$$

with respect to the Frobenius norm. Then by definition, there exists $\widetilde{A}_\Upsilon \in \mathcal{C}_\gamma(\epsilon, \lambda)$ such that $\|A_\Upsilon - \widetilde{A}_\Upsilon\|_{\mathrm{fro}} \le \epsilon^2$, which implies that

$$\big|w_z^\top A_\Upsilon w_z - w_z^\top \widetilde{A}_\Upsilon w_z\big| \le \|w_z\|_2^2 \cdot \|A_\Upsilon - \widetilde{A}_\Upsilon\|_{\mathrm{op}} \le \|A_\Upsilon - \widetilde{A}_\Upsilon\|_{\mathrm{fro}} \le \epsilon^2, \tag{J.24}$$

where we use the fact that

$$\|w_z\|_2^2 = \sum_{j=1}^{\gamma} |w_j|^2 = \sum_{j=1}^{\gamma} \sigma_j \cdot |\psi_j(z)|^2 = K(z,z) \le 1.$$

Thus, combining (J.23) and (J.24), and utilizing Corollary 4.2.13 in [69], we have

$$\log N_\infty(\epsilon, \mathcal{F}, \lambda) \le \log\big|\mathcal{C}_\gamma(\epsilon, \lambda)\big| \le \gamma^2 \cdot \log\big[1 + 8\sqrt{\gamma}/(\lambda \cdot \epsilon^2)\big] \le C_5 \cdot \gamma^2 \cdot \big[\log(1/\epsilon) + C_6\big],$$

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 the Cauchy-Schwarz inequality and Assumption 4.3, for any $z \in \mathcal{Z}$, by (J.22) we have

$$\begin{aligned}\big\|\phi(z) - \Pi_m\big[\phi(z)\big]\big\|_{\mathcal{H}} &= \Big\| \sum_{j=m+1}^{\infty} \sqrt{\sigma_j} \cdot \psi_j(z) \cdot \sqrt{\sigma_j} \cdot \psi_j \Big\|_{\mathcal{H}} = \Big\{ \sum_{j=m+1}^{\infty} \sigma_j \cdot [\psi_j(z)]^2 \Big\}^{1/2} \\ &\le \bigg(\sum_{j=m+1}^{\infty} \sigma_j \cdot \|\psi_j\|_\infty^2 \bigg)^{1/2} \le C_\psi \cdot \bigg(\sum_{j=m+1}^{\infty} \sigma_j^{1-2\tau} \bigg)^{1/2},\end{aligned} \tag{J.25}$$

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 a 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 (J.25) and the triangle inequality we have

$$\Big|\|\phi(z)\|_\Upsilon - \big\|\Pi_m\big[\phi(z)\big]\big\|_\Upsilon\Big| \leq \big\|\phi(z)-\Pi_m\big[\phi(z)\big]\big\|_\Upsilon \leq C_\psi/\sqrt{\lambda}\cdot\bigg(\sum_{j=m+1}^{\infty}\sigma_j^{1-2\tau}\bigg)^{1/2}. \tag{J.26}$$

Note that the eigenvalues $\{\sigma_j\}_{j\geq 0}$ admit γ-exponential decay under Assumption 4.3. We now upper bound the right-hand side of (J.26) by

$$\sup_{z\in\mathcal{Z}}\Big|\|\phi(z)\|_\Upsilon - \big\|\Pi_m\big[\phi(z)\big]\big\|_\Upsilon\Big| \leq C_\psi/\sqrt{\lambda}\cdot\bigg\{\sum_{j=m+1}^{\infty}C_1^{1-2\tau}\cdot\exp\big[-C_2\cdot(1-2\tau)\cdot j^\gamma\big]\bigg\}^{1/2}. \tag{J.27}$$

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 (J.27) and the monotonicity of $\exp(-u^\gamma)$, we further obtain

$$\sup_{z\in\mathcal{Z}}\Big|\|\phi(z)\|_\Upsilon - \big\|\Pi_m\big[\phi(z)\big]\big\|_\Upsilon\Big| \leq C_{3,\tau}\cdot\bigg[\int_m^\infty\exp(-C_{4,\tau}\cdot u^\gamma)\,\mathrm{d}u\bigg]^{1/2}. \tag{J.28}$$

Here we can take the supremum over $\mathcal{Z}$ because the right-hand side of (J.27) does not depend on z. Note that we have shown in (J.16) that

$$\int_m^\infty\exp(-C_{4,\tau}\cdot u^\gamma)\,\mathrm{d}u \leq \begin{cases} C_{4,\tau}^{-1}\cdot\exp(-C_{4,\tau}\cdot m^\gamma), & \text{if } \gamma\geq 1,\\ 2\cdot(\gamma\cdot C_{4,\tau})^{-1}\cdot\exp(-C_{4,\tau}\cdot m^\gamma)\cdot m^{1/\gamma-1}, & \text{if } \gamma\in(0,1),\end{cases} \tag{J.29}$$

where for the case of $\gamma\in(0,1)$, (J.29) holds for sufficient large m such that $m^\gamma\cdot C_{4,\tau}>2/\gamma-2$.

We now define m^* as the smallest integer such that

$$\int_{m^*}^\infty\exp(-C_{4,\tau}\cdot u^\gamma)\,\mathrm{d}u \leq \big[\epsilon/(2C_{3,\tau})\big]^2. \tag{J.30}$$

By (J.29), 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

$$m^*\leq C_{3,m}\cdot\big[\log(1/\epsilon)+C_{4,m}\big]^{1/\gamma}. \tag{J.31}$$

It is worth noting that the choice of m^* in (J.31) is uniform over all $z\in\mathcal{Z}$. Moreover, by (J.28), for such an m^*, it holds that

$$\sup_{z\in\mathcal{Z}}\Big|\|\phi(z)\|_\Upsilon - \big\|\Pi_{m^*}\big[\phi(z)\big]\big\|_\Upsilon\Big| \leq \epsilon/2. \tag{J.32}$$

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

$$[A_\Upsilon]_{j,k}=\big\langle\sqrt{\sigma_j}\cdot\psi_j,\sqrt{\sigma_k}\cdot\Upsilon\psi_k\big\rangle_{\mathcal{H}}. \tag{J.33}$$

Moreover, 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 (J.33) it holds that

$$\big\|\Pi_{m^*}\big[\phi(z)\big]\big\|_\Upsilon^2=\big\langle\Pi_{m^*}\big[\phi(z)\big],\Upsilon\Pi_{m^*}\big[\phi(z)\big]\big\rangle_{\mathcal{H}}=w_z^\top A_\Upsilon w_z. \tag{J.34}$$

Also, since $\|\Upsilon\|_{\mathrm{op}}\leq 1/\lambda$, the matrix operator norm of A_Υ is 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

$$\big|w_z^\top A_\Upsilon w_z - w_z^\top\widetilde{A}_\Upsilon w_z\big|\leq\|w_z\|_2^2\cdot\|A_\Upsilon-\widetilde{A}_\Upsilon\|_{\mathrm{op}}\leq\|A_\Upsilon-\widetilde{A}_\Upsilon\|_{\mathrm{fro}}\leq\epsilon^2/4. \tag{J.35}$$

Finally, for any $z\in\mathcal{Z}$, we define

$$f_\Upsilon(z)=w_z^\top\widetilde{A}_\Upsilon w_z=\sum_{j,k=1}^{m^*}\sqrt{\sigma_j\cdot\sigma_k}\cdot\psi_j(z)\cdot\psi_k(z)\cdot\big[\widetilde{A}_\Upsilon\big]_{jk}, \tag{J.36}$$

where $[\widetilde{A}_\Upsilon]_{jk}$ is the (j,k)-th entry of $\widetilde{A}_\Upsilon$ and m^* is specified in (J.30). We remark that $f_\Upsilon\colon \mathcal{Z} \to \mathbb{R}$ is well defined since m^* does not depend on z.

Finally, combining (J.32), (J.34), (J.35), and (J.36), we obtain

$$\begin{aligned}
\big\|\|\phi(z)\|_\Upsilon - f_\Upsilon\big\|_\infty &= \sup_{z\in\mathcal{Z}}\big|\|\phi(z)\|_\Upsilon - f_\Upsilon(z)\big| \\
&\le \sup_{z\in\mathcal{Z}}\Big|\|\phi(z)\|_\Upsilon - \big\|\Pi_{m^*}\big[\phi(z)\big]\big\|_\Upsilon\Big| + \sup_{z\in\mathcal{Z}}\Big|\big\|\Pi_{m^*}\big[\phi(z)\big]\big\|_\Upsilon - f_\Upsilon(z)\Big| \\
&\le \epsilon/2 + \sup_{z\in\mathcal{Z}}\Big|\sqrt{w_z^\top A_\Upsilon w_z} - \sqrt{w_z^\top \widetilde{A}_\Upsilon w_z}\Big| \le \epsilon/2 + \sup_{z\in\mathcal{Z}}\sqrt{\big|w_z^\top A_\Upsilon w_z - w_z^\top \widetilde{A}_\Upsilon w_z\big|} \le \epsilon.
\end{aligned}$$

This implies that $\{f_\Upsilon\colon \Upsilon \in \mathcal{C}_{m^*}(\epsilon,\lambda)\}$ forms an ϵ-cover of $\mathcal{F}(\lambda)$ in (I.12). Hence, we have that

$$N_\infty(\epsilon,\mathcal{F},\lambda) \le \big|\mathcal{C}_{m^*}(\epsilon/2,\lambda)\big|. \tag{J.37}$$

Furthermore, using Corollary 4.2.13 in [69], we have

$$\big|\mathcal{C}_{m^*}(\epsilon/2,\lambda)\big| \le \big[1 + 8\sqrt{m^*}/(\lambda\cdot\epsilon^2)\big]^{m^{*2}}. \tag{J.38}$$

Combining (J.31), (J.37), and (J.38), we finally have

$$\begin{aligned}
\log N_\infty(\epsilon,\mathcal{F},\lambda) &\le m^{*2}\cdot\log\big[1 + 8\sqrt{m^*}/(\lambda\cdot\epsilon^2)\big] \\
&\le C_{3,m}^2\cdot\big[\log(1/\epsilon) + C_{4,m}\big]^{2/\gamma}\cdot\log\Big\{1 + 8C_{3,m}^{1/2}\cdot\big[\log(1/\epsilon) + C_{4,m}\big]^{1/(2\gamma)}/(\lambda\cdot\epsilon^2)\Big\} \\
&\le C_5\cdot\big[\log(1/\epsilon) + C_6\big]^{1+2/\gamma},
\end{aligned}$$

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 and thus conclude the proof of the lemma. □

J.4 Technical Lemmas

Next, we present a few concentration inequalities. The first one provides concentration for standard self-normalized processes.

Lemma J.1 (Concentration of Self-Normalized Processes in RKHS [18])**.** 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 $\tau \ge 1$. 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-sub-Gaussian conditioning on $\mathcal{F}_{\tau-1}$:

$$\mathbb{E}[\epsilon_\tau|\mathcal{F}_{\tau-1}] = 0, \qquad \mathbb{E}[e^{\lambda\epsilon_\tau}|\mathcal{F}_{\tau-1}] \le e^{\lambda^2\sigma^2/2}, \qquad \forall\lambda\in\mathbb{R}.$$

Moreover, for any $t \ge 2$, 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 $t \ge 1$, we have

$$E_t^\top\big[(K_t + \eta\cdot I)^{-1} + I\big]^{-1}E_t \le \sigma^2\cdot\operatorname{logdet}\big[(1+\eta)\cdot I + K_t\big] + 2\sigma^2\cdot\log(1/\delta). \tag{J.39}$$

Moreover, if K_t is positive definite for all $t \ge 2$ with probability one, then the inequality in (J.39) also holds with $\eta = 0$.

Proof. See Theorem 1 in [18] for a detailed proof. □

Lemma J.2 (Lemma D.4 of [35])**.** Let $\{x_\tau\}_{\tau=1}^\infty$ and $\{\phi_\tau\}_{\tau=1}^\infty$ be $\mathcal{S}$-valued and $\mathcal{H}$-valued stochastic processes adapted to filtration $\{\mathcal{F}_\tau\}_{\tau=0}^\infty$, respectively, where we assume that $\|\phi_\tau\|_\mathcal{H} \le 1$ for all $\tau \ge 1$. Moreover, for any $t \ge 1$, 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

$$\sup_{V \in \mathcal{V}} \Big\| \sum_{\tau=1}^{t} \phi_\tau \big\{ V(x_\tau) - \mathbb{E}[V(x_\tau) | \mathcal{F}_{\tau-1}] \big\} \Big\|_{\Lambda_t^{-1}}^2 \tag{J.40}$$
$$\leq 2H^2 \cdot \operatorname{logdet}(I + K_t/\lambda) + 2H^2 t(\lambda - 1) + 4H^2 \log(\mathcal{N}_\epsilon/\delta) + 8t^2\epsilon^2/\lambda,$$

where $\mathcal{N}_\epsilon$ is the ϵ-covering number of $\mathcal{V}$ with respect to the distance $\operatorname{dist}(\cdot, \cdot)$.

Proof. Let $\mathcal{V}_\epsilon \subseteq \{V : \mathcal{S} \to [0, H]\}$ 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 $V' : \mathcal{S} \to \mathbb{R}$ in $\mathcal{N}_\epsilon$ such that $\operatorname{dist}(V, V') \leq \epsilon$. Let $\Delta_V = V - V'$. By the inequality $(a + b)^2 \leq 2a^2 + 2b^2$, we have

$$\Big\| \sum_{\tau=1}^{t} \phi_\tau \big\{ V(x_\tau) - \mathbb{E}[V(x_\tau) | \mathcal{F}_{\tau-1}] \big\} \Big\|_{\Lambda_t^{-1}}^2 \tag{J.41}$$
$$\leq 2 \cdot \Big\| \sum_{\tau=1}^{t} \phi_\tau \big\{ V'(x_\tau) - \mathbb{E}[V'(x_\tau) | \mathcal{F}_{\tau-1}] \big\} \Big\|_{\Lambda_t^{-1}}^2 + 2 \cdot \Big\| \sum_{\tau=1}^{t} \phi_\tau \big\{ \Delta_V(x_\tau) - \mathbb{E}[\Delta_V(x_\tau) | \mathcal{F}_{\tau-1}] \big\} \Big\|_{\Lambda_t^{-1}}^2.$$

To bound the first term on the right-hand side of (J.41), we apply Lemma J.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

$$\Big\| \sum_{\tau=1}^{t} \phi_\tau \big\{ \Delta_V(x_\tau) - \mathbb{E}[\Delta_V(x_\tau) | \mathcal{F}_{\tau-1}] \big\} \Big\|_{\Lambda_t^{-1}}^2 \leq t^2 \cdot (2\epsilon)^2/\lambda = 4t^2\epsilon^2/\lambda. \tag{J.42}$$

Thus, combining (J.41) and (J.42), we have

$$\sup_{V \in \mathcal{V}} \Big\| \sum_{\tau=1}^{t} \phi_\tau \big\{ V(x_\tau) - \mathbb{E}[V(x_\tau) | \mathcal{F}_{\tau-1}] \big\} \Big\|_{\Lambda_t^{-1}}^2$$
$$\leq \sup_{V' \in \mathcal{V}_\epsilon} 2 \cdot \Big\| \sum_{\tau=1}^{t} \phi_\tau \big\{ V'(x_\tau) - \mathbb{E}[V'(x_\tau) | \mathcal{F}_{\tau-1}] \big\} \Big\|_{\Lambda_t^{-1}}^2 + 8t^2\epsilon^2/\lambda. \tag{J.43}$$

Now we fix $V' \in \mathcal{V}_\epsilon$ 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 $\tau \geq 1$. We define an operator $\Phi : \mathcal{H} \to \mathbb{R}^t$ 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

$$\Big\| \sum_{\tau=1}^{t} \phi_\tau \big\{ V'(x_\tau) - \mathbb{E}[V'(x_\tau) | \mathcal{F}_{\tau-1}] \big\} \Big\|_{\Lambda_t^{-1}}^2 = \|\Phi_t^\top \varepsilon_t\|_{\Lambda_t^{-1}}^2 = \varepsilon_t^\top \Phi_t \Lambda_t^{-1} \Phi_t^\top \varepsilon_t$$
$$= \varepsilon_t^\top \Phi_t \Phi_t^\top (K_t + \lambda \cdot I)^{-1} \varepsilon_t = \varepsilon_t^\top K_t (K_t + \lambda \cdot I)^{-1} \varepsilon_t, \tag{J.44}$$

where the third inequality follows from (H.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

$$\varepsilon_t^\top K_t (K_t + \lambda \cdot I)^{-1} \varepsilon_t \leq \varepsilon_t^\top (K_t + \eta \cdot I)\big[I + (K_t + \eta \cdot I)\big]^{-1} \varepsilon_t$$
$$= \varepsilon_t^\top \big[(K_t + \eta \cdot I)^{-1} + I\big]^{-1} \varepsilon_t. \tag{J.45}$$

Notice that each entry of ε_t is bounded by H in absolute value since V' is bounded in $[0, H]$. By combining (J.43), (J.44), (J.45), Lemma J.1, and taking a union bound over $\mathcal{V}_\epsilon$, for any $\delta \in (0, 1)$, we obtain that, with probability at least $1 - \delta$,

$$\sup_{V' \in \mathcal{V}_\epsilon} \Big\| \sum_{\tau=1}^{t} \phi_\tau \big\{ V'(x_\tau) - \mathbb{E}[V'(x_\tau) | \mathcal{F}_{\tau-1}] \big\} \Big\|_{\Lambda_t^{-1}}^2$$
$$\leq H^2 \cdot \operatorname{logdet}[(1 + \eta) \cdot I + K_t] + 2H^2 \cdot \log(\mathcal{N}_\epsilon/\delta) \tag{J.46}$$

holds simultaneously for all $t \geq 1$. Moreover, notice that $(1 + \eta) \cdot I + K_t = [I + (1 + \eta)^{-1} \cdot K_t] \cdot [(1 + \eta) \cdot I]$, which implies that

$$\operatorname{logdet}\big[(1 + \eta) \cdot I + K_t\big] = \operatorname{logdet}\big[I + (1 + \eta)^{-1} \cdot K_t\big] + t \ln(1 + \eta)$$
$$\leq \operatorname{logdet}\big[I + (1 + \eta)^{-1} \cdot K_t\big] + \eta t. \tag{J.47}$$

Finally, combining (J.43), (J.46), and (J.47), we conclude that, simultaneously for all $t \geq 1$, (J.40) holds with probability at least $1-\delta$, which concludes the proof. □

Lemma J.3 ([1]). Let $\{\phi_t\}_{t\geq 1}$ 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 $\lambda \geq 1$ and $\mathcal{I}_{\mathcal{H}}$ is the identity mapping on $\mathcal{H}$. For any $t \geq 1$, 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

$$\sum_{j=1}^{t} \min\bigl\{1, \phi_j^\top \Lambda_{j-1}^{-1} \phi_j\bigr\} \leq 2\mathrm{logdet}(I + K_t/\lambda),$$

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

$$\mathrm{logdet}(I + K_t/\lambda) \leq \sum_{j=1}^{t} \phi_j^\top \Lambda_{j-1}^{-1} \phi_j \leq 2\mathrm{logdet}(I + K_t/\lambda).$$

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

$$\sum_{j=1}^{t} \min\bigl\{1, \phi_j^\top \Lambda_{j-1}^{-1} \phi_j\bigr\} \leq \sum_{j=1}^{t} 2\log\bigl(\min\bigl\{2, 1 + \phi_j^\top \Lambda_{j-1}^{-1} \phi_j\bigr\}\bigr) \leq 2\sum_{j=1}^{t} \log\bigl(1 + \phi_j^\top \Lambda_{j-1}^{-1} \phi_j\bigr). \tag{J.48}$$

Moreover, when additionally it is the case that $\sup_{j\geq 1} \|\phi_j\|_{\mathcal{H}} \leq 1$ for all $j \geq 0$, we have

$$\phi_j^\top \Lambda_{j-1}^{-1} \phi_j = \bigl\langle \phi_j, \Lambda_{j-1}^{-1} \phi_j \bigr\rangle_{\mathcal{H}} \leq \|\phi_j\|_{\mathcal{H}} \cdot \bigl\|\Lambda_{j-1}^{-1} \phi_j\bigr\|_{\mathcal{H}} \leq [\lambda_{\min}(\Lambda_0)]^{-1} \cdot \|\phi_j\|_{\mathcal{H}}^2 \leq 1. \tag{J.49}$$

Hence, applying the basic inequality $\log(1+x) \leq x \leq 2\log(1+x)$ to (J.49), we have

$$\sum_{j=1}^{t} \log\bigl(1 + \phi_j^\top \Lambda_{j-1}^{-1} \phi_j\bigr) \leq \sum_{j=1}^{t} \phi_j^\top \Lambda_{j-1}^{-1} \phi_j \leq 2\sum_{j=1}^{t} \log\bigl(1 + \phi_j^\top \Lambda_{j-1}^{-1} \phi_j\bigr). \tag{J.50}$$

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 eigenvalues 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

$$\Lambda_j = \Lambda_{j-1} + \phi_j \phi_j^\top = \Lambda_{j-1}^{1/2}\bigl(\mathcal{I}_{\mathcal{H}} + \Lambda_{j-1}^{-1/2} \phi_j \phi_j^\top \Lambda_{j-1}^{-1/2}\bigr)\Lambda_{j-1}^{1/2},$$

which implies that

$$\begin{aligned} \mathrm{logdet}(\Lambda_j) &= \mathrm{logdet}(\Lambda_{j-1}) + \mathrm{logdet}\bigl(\mathcal{I}_{\mathcal{H}} + \Lambda_{j-1}^{-1/2} \phi_j \phi_j^\top \Lambda_{j-1}^{-1/2}\bigr) \\ &= \mathrm{logdet}\bigl(\Lambda_{j-1}\bigr) + \mathrm{logdet}\bigl(1 + \phi_j^\top \Lambda_{j-1}^{-1} \phi_j\bigr) \end{aligned} \tag{J.51}$$

Moreover, by direct computation, for any $t \geq 1$, we have

$$\det(\Lambda_t \Lambda_0^{-1}) = \det(I + K_t/\lambda). \tag{J.52}$$

Hence, combining (J.51), and (J.52), we obtain that

$$\sum_{j=1}^{t} \log\bigl(1 + \phi_j^\top \Lambda_{j-1}^{-1} \phi_j\bigr) = \mathrm{logdet}(\Lambda_t \Lambda_0^{-1}) = \mathrm{logdet}(I + K_t/\lambda). \tag{J.53}$$

Finally, combining (J.48), (J.50) and (J.53), we conclude the proof of this lemma. □