Optimal Rates of Kernel Ridge Regression under Source Condition in Large Dimensions

Haobo Zhang

ZHANG-HB21@MAILS.TSINGHUA.EDU.CN

Yicheng Li

LIYC22@MAILS.TSINGHUA.EDU.CN

Weihao Lu

LUWH19@MAILS.TSINGHUA.EDU.CN

Qian Lin *

QIANLIN@TSINGHUA.EDU.CN

Center for Statistical Science, Department of Industrial Engineering Tsinghua University

Abstract

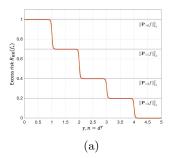
Motivated by the studies of neural networks (e.g.,the neural tangent kernel theory), we perform a study on the large-dimensional behavior of kernel ridge regression (KRR) where the sample size $n \asymp d^{\gamma}$ for some $\gamma > 0$. Given an RKHS \mathcal{H} associated with an inner product kernel defined on the sphere \mathbb{S}^d , we suppose that the true function $f_{\rho}^* \in [\mathcal{H}]^s$, the interpolation space of \mathcal{H} with source condition s > 0. We first determined the exact order (both upper and lower bound) of the generalization error of kernel ridge regression for the optimally chosen regularization parameter λ . We then further showed that when $0 < s \le 1$, KRR is minimax optimal; and when s > 1, KRR is not minimax optimal (a.k.a. the saturation effect). Our results illustrate that the curves of rate varying along γ exhibit the periodic plateau behavior and the multiple descent behavior and show how the curves evolve with s > 0. Interestingly, our work provides a unified viewpoint of several recent works on kernel regression in the large-dimensional setting, which correspond to s = 0 and s = 1 respectively.

Keywords: kernel methods, high-dimensional statistics, reproducing kernel Hilbert space, minimax optimality, saturation effect

1. Introduction

The recent studies of neural network theory have brought the renaissance of kernel methods, since the neural tangent kernel (Jacot et al., 2018) provides a natural surrogate to understand the wide neural network (Arora et al., 2019; Lee et al., 2019; Lai et al., 2023). When the dimension of data is fixed, there has been extensive literature studying the generalization behavior of kernel ridge regression (KRR), one of the most popular kernel methods, e.g., Caponnetto and de Vito (2007); Fischer and Steinwart (2020); Cui et al. (2021), etc.. Researchers usually use two crucial factors to characterize KRR's generalization behavior: capacity condition and source condition. Supposing eigenvalues associated with the RKHS \mathcal{H} are $\{\lambda_i\}_{i=1}^{\infty}$, the capacity condition (also known as effective dimension) assumes $\mathcal{N}_1(\lambda) := \sum_{i=1}^{\infty} \lambda_i/(\lambda_i + \lambda) \approx \lambda^{-\frac{1}{\beta}}$ for some $\beta > 1$, where λ will represent the regularization parameter

^{*.} Corresponding author



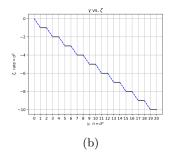


Figure 1: Left: The curve of generalization error for estimating $f_{\rho}^* \in L^2$ (Figure 5 in Ghorbani et al. (2021)). Right: The curve of the minimax rates for estimating $f_{\rho}^* \in \mathcal{H}$ (Figure 2(b) in Lu et al. (2023)).

in KRR. The capacity condition characterizes the size of \mathcal{H} and is frequently stated as an equivalent eigenvalue decay condition: $\lambda_i \approx i^{-\beta}, \beta > 1$. The source condition assumes that the true function f_{ρ}^* falls into $[\mathcal{H}]^s$, an interpolation space of \mathcal{H} for some s > 0. It characterizes the relative smoothness of f_{ρ}^* with respect to \mathcal{H} : the larger s is, the "smoother" f_{ρ}^* is and the easier it can be estimated. Under this framework, many interesting topics about KRR's generalization behavior were studied. For instance, the minimax optimality of KRR (Fischer and Steinwart, 2020; Zhang et al., 2023b) when $0 < s \le 2$, the saturation effect of KRR (Bauer et al., 2007; Li et al., 2023b) when s > 2, the generalization ability of kernel interpolation (Beaglehole et al., 2023; Li et al., 2023a) and the learning curve of KRR (Cui et al., 2021; Li et al., 2023c), etc. We refer to Section 1.1 for more related work about these topics. These results help us clarify several puzzle points. For example, Li et al. (2023b) implies that when the true function is smooth enough (e.g., s > 2), the early stopping kernel gradient flow is often better than the kernel ridge regression and Li et al. (2023a) asserts that if a wide neural network overfits the data, it generalizes poorly.

Since neural networks often perform well on data with large dimensionality where $n \simeq d^{\gamma}$ for some $\gamma > 0$, we expect that the studies of kernel regression in large-dimensional data can provide us more guidance about the generalization behavior of neural network for large-dimensional data. However, in contrast to the rich theoretical results about kernel regression in the fixed-dimensional setting, much less is known about the aforementioned topics in the large-dimensional setting. Since Karoui (2010) provided an approximation of the kernel random matrix when $n \simeq d$, few works have been done for kernel regression in large-dimensional or high-dimensional settings until recently. The first obstacle is that if d is large, the eigenvalues of the RKHS usually depend on d in an unpleasant way. Therefore, the polynomial eigenvalue decay rate assumption $\lambda_i \simeq i^{-\beta}$ must not be true (e.g., the inner product kernel on the sphere in Section 3.2). Second, we find that $\mathcal{N}_1(\lambda)$ is not enough to characterize a tight upper bound of the convergence rate, which is a significant difference from the fixed-dimensional setting. We will see that more information about the eigenvalues (RKHS) is needed, for instance, $\mathcal{N}_2(\lambda) := \sum_{i=1}^{\infty} (\lambda_i/(\lambda_i + \lambda))^2$ which will be introduced in (4).

There are several recent works investigating the generalization error of kernel regression in the large-dimensional setting where $n \approx d^{\gamma}$, $\gamma > 0$. Ghorbani et al. (2021) considers the square-integrable function space on the sphere \mathbb{S}^d and proves that when γ is a non-integer, KRR is consistent if and only if the true function is a polynomial with a fixed degree $< \gamma$. They also qualitatively reveal that the excess risk exhibits a periodic plateau behavior (Figure 1(a)). Liu et al. (2021) considers the setting $n \approx d$ and assumes source condition to be $s \in (0,2]$. They give an upper bound of the generalization error with respect to bias and variance. Using this upper bound, they demonstrate that there could be multiple shapes of the generalization curve as the sample size increases. A more recent work Lu et al. (2023) studies early stopping kernel gradient flow in the large-dimensional setting $n \simeq d^{\gamma}$. Assuming $f_{\rho}^* \in \mathcal{H}$ and considering the inner product kernel on the sphere \mathbb{S}^d , they prove an upper bound of the convergence rate and show the minimax optimality of early stopping kernel gradient flow. Interestingly, their results indicate that the minimax rate of the kernel regression for $f_{\rho}^* \in \mathcal{H}$ exhibits the similar periodic plateau phenomenon (Figure 1(b)). This raises a natural and interesting question: is there a unified way to explain the periodic plateau behavior appeared in Lu et al. (2023) and Ghorbani et al. (2021)?

Suppose that \mathcal{H} is an RKHS associated with an inner product kernel defined on \mathbb{S}^d . The main focus of this paper aims to derive the matching upper and lower bounds of the generalization error and discuss the minimax optimality of KRR for general source condition s > 0, i.e., the true regression function $f_{\rho}^* \in [\mathcal{H}]^s$. Allowing s to vary is not only a more reasonable assumption for the true function, but also provides us a natural framework to clarify the relation between the results in Lu et al. (2023) and Ghorbani et al. (2020). In fact, a little bit interpolation space theory suggests that both the results in Ghorbani et al. (2021) and Lu et al. (2023) are two special cases of our results corresponding to s = 0 and s = 1 respectively. This paper has the following contributions:

- We consider a more general framework than the traditional capacity-source condition. We introduce $\mathcal{N}_1(\lambda), \mathcal{N}_2(\lambda), \mathcal{M}_1(\lambda)$ and $\mathcal{M}_2(\lambda)$ in (4), which are key quantities depending on the RKHS, the true function and the regularization parameter λ . Under mild assumptions, we use these key quantities to express the matching upper and lower bounds of the generalization error as long as the regularization parameter satisfies some approximation conditions (Theorem 1). This framework makes few assumptions on the eigenvalues of the RKHS and the true function, thus enabling us to handle the large-dimensional setting and general source condition later. In the fixed-dimensional setting, our results in Theorem 1 also recovers the state-of-the-art theoretical results about the exact convergence rates of KRR in Li et al. (2023c).
- We then add source condition into our new framework and consider the inner product kernel on the sphere \mathbb{S}^d . When $n \asymp d^{\gamma}$, we derive exact convergence rates (both upper and lower bounds) of the generalization error under the best choice of regularization parameter for any source condition s > 0 and almost all $\gamma > 0$ (Theorem 2 for $s \ge 1$ and Theorem 3 for 0 < s < 1). We will see that the curves of rate varying along γ show similar periodic plateau and multiple descent behavior as in Lu et al. (2023). Moreover, we will see that the shapes of curves vary with s and are totally different when 0 < s < 1 and $s \ge 1$, with even more intriguing results in the limiting case $s \to 0$ and $s \to 2$.

• For the inner product kernel on the sphere \mathbb{S}^d , we further derive the corresponding minimax lower bound for all s>0 and $\gamma>0$. When 0< s<1, the exact rates in Theorem 3 match the minimax lower bound, and thus we prove the minimax optimality of KRR. When s>1, the KRR is not minimax optimal, i.e., we discover a new version of the saturation effect of KRR. In the fixed-dimensional setting, the saturation effect of KRR only happens when s>2. In the large-dimensional setting, we prove that a similar phenomenon also happens for $1< s\leq 2$. Specifically, for any s>1, there will be corresponding ranges of γ such that the convergence rates of KRR can not achieve the minimax lower bound even under the best choice of regularization parameter.

1.1 Related work

In the introduction, we have mentioned several interesting topics about the KRR's generalization behavior. These topics have been well-studied in the fixed-dimensional setting. The first essential question is the minimax optimality of KRR. Under the framework of capacity condition and source condition (s), Caponnetto and de Vito (2007) proves the minimax optimality of KRR when $1 \le s \le 2$. Then, extensive literature (see, e.g., Steinwart et al. 2009; Lin et al. 2018; Fischer and Steinwart 2020; Zhang et al. 2023a,b and the reference therein) studies the mis-specified case (0 < s < 1), where Zhang et al. (2023a) proves the minimax optimality for all $0 < s \le 2$ under further embedding index condition. The second question is the saturation effect of KRR when s > 2, which is conjectured by Bauer et al. (2007); Gerfo et al. (2008) and rigorously proved by Li et al. (2023b). Thirdly, due to the fantastic performance of overparameterized neural networks, the generalization ability of kernel interpolation has also raised a lot of interest. The results in Rakhlin and Zhai (2019); Buchholz (2022); Beaglehole et al. (2023); Li et al. (2023a) imply that kernel interpolation can not generalize in fixed-dimensional setting. Last but not least, Bordelon et al. (2020); Cui et al. (2021); Li et al. (2023c) study the learning curve of KRR, i.e., the exact generalization error (or exact order) for any regularization parameter $\lambda > 0$.

In the large-dimensional setting, the answers to the above questions are not clear yet. Many researchers have studied these problems from different angles and settings. A line of work uses the tools of high-dimensional kernel random matrix approximation from Karoui (2010) and studies the generalization ability of kernel interpolation (Liang and Rakhlin, 2020; Liang et al., 2020). When $n \approx d$, Liang and Rakhlin (2020) gives an upper bound of the generalization error of kernel interpolation and claims that the upper bound tends to 0 when the data exhibits a low-dimensional structure. Further, when $n \approx d^{\gamma}$, $\gamma > 0$, Liang et al. (2020) gives an upper bound with a specific convergence rate, which implies that kernel interpolation can generalize if and only if γ is not an integer. One closely related topic is the benign overfitting phenomenon, which we refer to Bartlett et al. (2020); Hastie et al. (2022); Muthukumar et al. (2020); Tsigler and Bartlett (2023).

Another line of work follows Ghorbani et al. (2021), which has been mentioned in the introduction. This line of work adopts the square-integrable assumption of the true function and aims to obtain the exact generalization error of kernel methods in various settings (Ghorbani et al., 2020; Mei et al., 2022; Mei and Montanari, 2022; Ghosh et al., 2021; Xiao et al., 2022; Hu and Lu, 2022; Misiakiewicz, 2022; Donhauser et al., 2021). To our knowledge, Lu et al. (2023) is the only literature that provides the minimax optimality result for specific

kernel methods. As discussed in the introduction, Lu et al. (2023) considers the s = 1 case $(f_{\rho}^* \in \mathcal{H})$ and kernel early stopping gradient flow. We will provide a detailed discussion on Lu et al. (2023) and Ghorbani et al. (2021) in Section 4. If we follow the line of research in the fixed-dimensional setting, considering general source condition s > 0 and studying the minimax optimality (saturation effect) of KRR are essential steps to understand KRR's generalization behavior in the large-dimensional setting.

2. Preliminaries

Let a compact set $\mathcal{X} \subseteq \mathbb{R}^d$ be the input space and $\mathcal{Y} \subseteq \mathbb{R}$ be the output space. Let $\rho = \rho_d$ be an unknown probability distribution on $\mathcal{X} \times \mathcal{Y}$ satisfying $\int_{\mathcal{X} \times \mathcal{Y}} y^2 d\rho(\mathbf{x}, y) < \infty$ and denote the corresponding marginal distribution on \mathcal{X} as $\mu = \mu_d$. We use $L^p(\mathcal{X}, \mu)$ (in short L^p) to represent the L^p -spaces. Throughout the paper, we make the following assumption:

Assumption 1 Suppose that $\mathcal{H} = \mathcal{H}_d$ is a separable RKHS on $\mathcal{X} \subset \mathbb{R}^d$ with respect to a continuous kernel function $k = k_d$ satisfying

$$\sup_{\boldsymbol{x}\in\mathcal{X}}k_d(\boldsymbol{x},\boldsymbol{x})\leq\kappa^2,$$

where κ is an absolute constant.

Since we allow the dimension d to diverge to infinity as $n \to \infty$, the RKHS may vary with d and we suppose that Assumption 1 holds uniformly for all d. For brevity of notations, we frequently omit the index d in the rest of this paper.

Suppose that the samples $\{(x_i, y_i)\}_{i=1}^n$ are i.i.d. sampled from ρ . Kernel ridge regression (KRR) constructs an estimator \hat{f}_{λ} by solving the penalized least square problem

$$\hat{f}_{\lambda} = \operatorname*{arg\,min}_{f \in \mathcal{H}} \left(\frac{1}{n} \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i))^2 + \lambda ||f||_{\mathcal{H}}^2 \right),$$

where $\lambda > 0$ is referred to as the regularization parameter.

Denote the samples as $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$ and $\boldsymbol{y} = (y_1, \dots, y_n)^{\top}$. The representer theorem (see, e.g., Steinwart and Christmann 2008) gives an explicit expression of the KRR estimator, i.e.,

$$\hat{f}_{\lambda}(\mathbf{x}) = \mathbb{K}(\mathbf{x}, \mathbf{X})(\mathbb{K}(\mathbf{X}, \mathbf{X}) + n\lambda \mathbf{I})^{-1}\mathbf{y}, \tag{1}$$

where

$$\mathbb{K}(\boldsymbol{X}, \boldsymbol{X}) = (k(\boldsymbol{x}_i, \boldsymbol{x}_j))_{n \times n}, \quad \mathbb{K}(\boldsymbol{x}, \boldsymbol{X}) = (k(\boldsymbol{x}, \boldsymbol{x}_1), \dots, k(\boldsymbol{x}, \boldsymbol{x}_n)).$$

Denote the conditional mean:

$$f_{
ho}^*(oldsymbol{x}) = f_{
ho,d}^*(oldsymbol{x}) \coloneqq \mathbb{E}_{
ho}[\ y \mid oldsymbol{x}\] = \int_{\mathcal{Y}} y \ \mathrm{d}
ho(y | oldsymbol{x}).$$

We are interested in the convergence rates of the generalization error (excess risk) of \hat{f}_{λ} :

$$\mathbb{E}_{\boldsymbol{x} \sim \mu} \left[\left(\hat{f}_{\lambda}(\boldsymbol{x}) - f_{\rho}^{*}(\boldsymbol{x}) \right)^{2} \right] = \left\| \hat{f}_{\lambda} - f_{\rho}^{*} \right\|_{L^{2}}^{2}.$$

Notations. We use asymptotic notations $O(\cdot)$, $o(\cdot)$, $o(\cdot)$, $O(\cdot)$ and $O(\cdot)$. We also write $a_n \approx b_n$ for $a_n = O(b_n)$; $a_n \lesssim b_n$ for $a_n = O(b_n)$; $a_n \lesssim b_n$ for $a_n = O(b_n)$; $a_n \ll b_n$ for $a_n = o(b_n)$. We will also use the probability versions of the asymptotic notations such as $O_{\mathbb{P}}(\cdot)$, $O_{\mathbb{P}}(\cdot)$, $O_{\mathbb{P}}(\cdot)$, $O_{\mathbb{P}}(\cdot)$. For instance, we say the random variables X_n, Y_n satisfying $X_n = O_{\mathbb{P}}(Y_n)$ if and only if for any $\epsilon > 0$, there exist a constant C_{ϵ} and N_{ϵ} such that $O(|X_n| \geq C_{\epsilon}|Y_n|) \leq \epsilon, \forall n > N_{\epsilon}$.

2.1 Integral operator and interpolation space

Denote the natural embedding inclusion operator as $S_k : \mathcal{H} \to L^2(\mathcal{X}, \mu)$. Then its adjoint operator $S_k^* : L^2(\mathcal{X}, \mu) \to \mathcal{H}$ is an integral operator, i.e., for $f \in L^2(\mathcal{X}, \mu)$ and $\mathbf{x} \in \mathcal{X}$, we have

$$(S_k^* f)(\mathbf{x}) = \int_{\mathcal{X}} k(\mathbf{x}, \mathbf{x}') f(\mathbf{x}') d\mu(\mathbf{x}').$$

Under Assumption 1, S_k and S_k^* are Hilbert-Schmidt operators (thus compact) and the HS norms (denoted as $\|\cdot\|_2$) satisfy that

$$\|S_k^*\|_2 = \|S_k\|_2 = \|k\|_{L^2(\mathcal{X},\mu)} := \left(\int_{\mathcal{X}} k(\boldsymbol{x}, \boldsymbol{x}) d\mu(\boldsymbol{x})\right)^{1/2} \le \kappa.$$

Next, we can define two integral operators:

$$L_k := S_k S_k^* : L^2(\mathcal{X}, \mu) \to L^2(\mathcal{X}, \mu), \qquad T := S_k^* S_k : \mathcal{H} \to \mathcal{H}. \tag{2}$$

 L_k and T are self-adjoint, positive-definite and trace class (thus Hilbert-Schmidt and compact) and the trace norms (denoted as $\|\cdot\|_1$) satisfy that

$$||L_k||_1 = ||T||_1 = ||S_k||_2^2 = ||S_k^*||_2^2$$
.

The spectral theorem for self-adjoint compact operators yields that there is an at most countable index set N, a non-increasing summable sequence $\{\lambda_i\}_{i\in N}\subseteq (0,\infty)$ and a family $\{e_i\}_{i\in N}\subseteq \mathcal{H}$, such that $\{e_i\}_{i\in N}$ is an orthonormal basis (ONB) of $\overline{\operatorname{ran} S_k}\subseteq L^2(\mathcal{X},\mu)$ and $\{\lambda_i^{1/2}e_i\}_{i\in N}$ is an ONB of \mathcal{H} . Further, the integral operators can be written as

$$L_k = \sum_{i \in N} \lambda_i \langle \cdot, e_i \rangle_{L^2} e_i$$
 and $T = \sum_{i \in N} \lambda_i \langle \cdot, \lambda_i^{1/2} e_i \rangle_{\mathcal{H}} \lambda_i^{1/2} e_i$.

We refer to $\{e_i\}_{i\in N}$ and $\{\lambda_i\}_{i\in N}$ as the eigenfunctions and eigenvalues. The celebrated Mercer's theorem (see, e.g., Steinwart and Christmann 2008, Theorem 4.49) shows that

$$k\left(\boldsymbol{x}, \boldsymbol{x}'\right) = \sum_{i \in N} \lambda_{i} e_{i}(\boldsymbol{x}) e_{i}\left(\boldsymbol{x}'\right), \quad \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X},$$

where the convergence is absolute and uniform for x, x'.

Since we are going to consider the source condition in this paper, we need to introduce the interpolation spaces (power spaces) of RKHS. For any $s \geq 0$, the fractional power integral operator $L_k^s: L^2(\mathcal{X}, \mu) \to L^2(\mathcal{X}, \mu)$ is defined as

$$L_k^s(f) = \sum_{i \in N} \lambda_i^s \, \langle f, e_i \rangle_{L^2} \, e_i.$$

Then the interpolation space (power space) $[\mathcal{H}]^s$ is defined as

$$[\mathcal{H}]^s := \operatorname{Ran} L_k^{s/2} = \left\{ \sum_{i \in N} a_i \lambda_i^{s/2} e_i : (a_i)_{i \in N} \in \ell_2(N) \right\} \subseteq L^2(\mathcal{X}, \mu), \tag{3}$$

equipped with the inner product

$$\langle f, g \rangle_{[\mathcal{H}]^s} = \left\langle L_k^{-\frac{s}{2}} f, L_k^{-\frac{s}{2}} g \right\rangle_{L^2}.$$

It is easy to show that $[\mathcal{H}]^s$ is also a separable Hilbert space with orthogonal basis $\{\lambda_i^{s/2}e_i\}_{i\in N}$. Specially, we have $[\mathcal{H}]^0\subseteq L^2(\mathcal{X},\mu)$ and $[\mathcal{H}]^1=\mathcal{H}$. For $0< s_1< s_2$, the embeddings $[\mathcal{H}]^{s_2}\hookrightarrow [\mathcal{H}]^{s_1}\hookrightarrow [\mathcal{H}]^0$ exist and are compact (Fischer and Steinwart, 2020). For the functions in $[\mathcal{H}]^s$ with larger s, we say they have higher regularity (smoothness) with respect to the RKHS.

In the following of this paper, we assume $|N| = \infty$. Also note that $\{\lambda_i\}_{i=1}^{\infty}$ and $\{e_i\}_{i=1}^{\infty}$ are dependent on \mathcal{H} , thus are dependent on d.

3. Main results

3.1 KRR's generalization error in the general case

In this subsection, we consider a general framework for studying the generalization error of KRR, where we make quite mild assumptions on the RKHS \mathcal{H} and the true function f_{ρ}^* . Note that in this framework, we allow d to diverge to infinity as sample size n and allow \mathcal{H}, f_{ρ}^* to change with d. Therefore, the results in this subsection are applicable in the large-dimensional setting $n \approx d^{\gamma}, \gamma > 0$.

Given the RKHS \mathcal{H} and denote the true function as $f_{\rho}^* = \sum_{i=1}^{\infty} f_i e_i(\boldsymbol{x}) \in L^2(\mathcal{X}, \mu)$. We define the following important quantities:

$$\mathcal{N}_{1}(\lambda) = \sum_{i=1}^{\infty} \left(\frac{\lambda_{i}}{\lambda_{i} + \lambda}\right); \quad \mathcal{N}_{2}(\lambda) = \sum_{i=1}^{\infty} \left(\frac{\lambda_{i}}{\lambda_{i} + \lambda}\right)^{2};$$

$$\mathcal{M}_{1}(\lambda) = \operatorname{ess sup}_{\boldsymbol{x} \in \mathcal{X}} \left| \sum_{i=1}^{\infty} \left(\frac{\lambda}{\lambda_{i} + \lambda} f_{i} e_{i}(\boldsymbol{x})\right) \right|; \quad \mathcal{M}_{2}(\lambda) = \sum_{i=1}^{\infty} \left(\frac{\lambda}{\lambda_{i} + \lambda} f_{i}\right)^{2}.$$
(4)

Assumption 2 Suppose that for some absolute constant $\sigma > 0$,

$$\mathbb{E}_{(\boldsymbol{x},y)\sim\rho}\left[\left(y-f_{\rho}^{*}(\boldsymbol{x})\right)^{2}\;\middle|\; \boldsymbol{x}\right]=\sigma^{2},\quad \mu\text{-}a.e.\;\; \boldsymbol{x}\in\mathcal{X}.$$

Assumption 2 assumes that the noise is non-vanishing and it holds for common nonparametric regression model $y = f_{\rho}^*(\mathbf{x}) + \epsilon$ where ϵ is an independent non-zero noise.

Assumption 3 Suppose that

$$\operatorname{ess sup}_{\boldsymbol{x} \in \mathcal{X}} \sum_{i=1}^{\infty} \left(\frac{\lambda_i}{\lambda_i + \lambda} \right)^2 e_i^2(\boldsymbol{x}) \le \mathcal{N}_2(\lambda), \tag{5}$$

and

$$\operatorname{ess sup}_{\boldsymbol{x} \in \mathcal{X}} \sum_{i=1}^{\infty} \frac{\lambda_i}{\lambda_i + \lambda} e_i^2(\boldsymbol{x}) \le \mathcal{N}_1(\lambda). \tag{6}$$

In fact, it is allowed to multiply an absolute constant C on the right sides of (5) and (6). Without loss of generality, we consider the constant to be C = 1. Assumption 3 naturally holds for RKHSs with uniformly bounded eigenfunctions, i.e., $\sup_{i\geq 1} \sup_{\boldsymbol{x}\in\mathcal{X}} |e_i(\boldsymbol{x})| \leq 1$. One can also show that RKHSs associated with inner product kernel on the sphere with uniform distribution satisfy Assumption 3 (see Lemma 20).

Now we begin to state the first important theorem in this paper.

Theorem 1 Let $\mathcal{N}_1, \mathcal{N}_2, \mathcal{M}_1, \mathcal{M}_2$ be defined as (4), and let d = d(n) which is allowed to diverge with $n \to \infty$. Suppose that Assumption 1, 2 and 3 hold. Let \hat{f}_{λ} be the KRR estimator defined by (1). If the following approximation conditions hold for some $\lambda = \lambda(d, n) \to 0$:

$$\frac{\mathcal{N}_1(\lambda)}{n} \ln n = o(1); \quad n^{-1} \mathcal{N}_1(\lambda)^2 \ln n = o\left(\mathcal{N}_2(\lambda)\right); \quad n^{-1} \mathcal{N}_1(\lambda)^{\frac{1}{2}} \mathcal{M}_1(\lambda) = o\left(\mathcal{M}_2(\lambda)^{\frac{1}{2}}\right), \quad (7)$$

then we have

$$\mathbb{E}\left[\left\|\hat{f}_{\lambda} - f_{\rho}^{*}\right\|_{L^{2}}^{2} \mid \boldsymbol{X}\right] = \Theta_{\mathbb{P}}\left(\frac{\sigma^{2}\mathcal{N}_{2}(\lambda)}{n} + \mathcal{M}_{2}(\lambda)\right). \tag{8}$$

The notation $\Theta_{\mathbb{P}}$ only involves absolute constants.

Note that the notation $o(\cdot)$ represents the limit as $n \to \infty$ and we allow d = d(n) to diverge to infinity with n. Theorem 1 provides the matching upper and lower bounds (8) for all λ satisfying the approximation conditions (7). Generally speaking, the conditions in (7) are more likely to hold for larger λ . For instance, we will show in the proof of Theorem 2 that if the conditions in (7) hold for $\lambda_0 = d^{-l_0}$ for some $l_0 > 0$, then they hold for all $\lambda = d^{-l}$, $0 < l < l_0$.

Basically, in (8), the term $\sigma^2 \mathcal{N}_2(\lambda)/n$ corresponds to the variance and $\mathcal{M}_2(\lambda)$ corresponds to the bias. An obvious relation is $\mathcal{N}_2(\lambda) \leq \mathcal{N}_1(\lambda)$, thus the first approximation condition in (7) guarantees that $\lambda = \lambda(d, n)$ is not that small and the variance term tends to 0. In addition, if we take the traditional source condition assumption, i.e., $\|f_{\rho}^*\|_{[\mathcal{H}]^s} \leq R$ for some constant R and s > 0, easy calculation shows that $\mathcal{M}_2(\lambda) \leq C\lambda^{\min\{s,2\}}$ for some constant C only depending on R and the constant κ in Assumption 1. This implies that the bias term tends to 0 as $\lambda = \lambda(d, n) \to 0$.

Under the capacity-source condition framework in the fixed-dimensional setting (as discussed in the introduction), Theorem 1 also recovers the state-of-the-art results in Li et al. (2023c). We emphasize that proving such tight bounds of the generalization error in the large-dimensional setting is nontrivial. In addition, the original bounds of the key quantities in (4) under the capacity-source condition framework are no longer sufficient in the large-dimensional setting. Given the information about the kernel and the true function, detailed calculations of $\mathcal{N}_1(\lambda), \mathcal{N}_2(\lambda), \mathcal{M}_1(\lambda)$ and $\mathcal{M}_2(\lambda)$ will be needed (see, e.g., Appendix B.2 for inner product kernel on the sphere).

3.2 Applications to inner product kernel on the sphere

In this subsection, we consider the inner product kernel on the sphere with uniform distribution. In the large-dimensional setting $n \approx d^{\gamma}$, $\gamma > 0$ and under further source condition assumption, we apply Theorem 1 to prove the exact convergence rates of the generalization error of the KRR estimator. Then, we derive the corresponding minimax lower bound, which enables us to discuss the minimax optimality and the saturation effect of KRR.

Suppose that $\mathcal{X} = \mathbb{S}^d$ and μ is the uniform distribution on \mathbb{S}^d . We consider the inner product kernel, i.e., there exists a function $\Phi(t) : [-1,1] \to \mathbb{R}$ such that $k_d(\boldsymbol{x}, \boldsymbol{x}') = \Phi(\langle \boldsymbol{x}, \boldsymbol{x}' \rangle), \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{S}^d$. Then Mercer's decomposition for the inner product kernel is given in the basis of spherical harmonics:

$$k_d(\boldsymbol{x}, \boldsymbol{x'}) = \sum_{k=0}^{\infty} \mu_k \sum_{l=1}^{N(d,k)} Y_{k,l}(\boldsymbol{x}) Y_{k,l}(\boldsymbol{x'}),$$

where $\{Y_{k,l}\}_{l=1}^{N(d,k)}$ are spherical harmonic polynomials of degree k; μ_k are the eigenvalues with multiplicity N(d,0)=1; $N(d,k)=\frac{2k+d-1}{k}\cdot\frac{(k+d-2)!}{(d-1)!(k-1)!}, k=1,2,\cdots$.

Assumption 4 (Inner product kernel) Suppose that $k = \{k_d\}_{d=1}^{\infty}$ satisfies

$$k_d(\boldsymbol{x}, \boldsymbol{x}') = \Phi\left(\langle \boldsymbol{x}, \boldsymbol{x}' \rangle\right), \ \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{S}^d,$$

where $\Phi(t) \in \mathcal{C}^{\infty}([-1,1])$ is a fixed function independent of d and

$$\Phi(t) = \sum_{j=0}^{\infty} a_j t^j, \ a_j > 0, \ \forall j = 0, 1, 2, \dots$$

Assumption 4 formally defines the kernel considered in this subsection. The purpose of assuming all the coefficients a_j to be positive is to keep the main results and proofs clean. In fact, the proof is similar for other inner product kernels as long as we know which coefficients are positive, for instance, the neural tangent kernel in the following subsection. We assume $\Phi(t)$ (or $\{a_j\}_{j=0}^{\infty}$) to be fixed and we will ignore the dependence of constants on it in the rest of our paper.

The inner product kernel has attracted a lot of research (Liang et al., 2020; Ghorbani et al., 2021; Misiakiewicz, 2022; Xiao et al., 2022; Lu et al., 2023, etc.) and we have a concise characterization of μ_k and N(d, k), which enables us to calculate the exact convergence rates of the key quantities in (4). We refer to Lemma 17, 18 and 19 in Appendix B.1 for details about μ_k and N(d, k). The extension to general kernel can be extremely complicated and existing results also only consider the case where \mathcal{X} is the sphere (as this paper) or discrete hypercube (see, e.g., Mei et al. 2022; Aerni et al. 2022).

In the next assumption, we formally introduce the source condition, which characterizes the relative smoothness of f_{ρ}^* with respect to \mathcal{H} .

Assumption 5 (Source condition)

(a) Suppose that $f_{\rho}^*(\mathbf{x}) = f_{\rho,d}^*(\mathbf{x}) = \sum_{i=1}^{\infty} f_i e_i(\mathbf{x}) \in [\mathcal{H}]^s$ for some s > 0 and satisfies that,

$$\left\| f_{\rho}^* \right\|_{[\mathcal{H}]^s} \le R_{\gamma},\tag{9}$$

where R_{γ} is a constant only depending on γ .

(b) Denote q as the smallest integer such that $q > \gamma$ and $\mu_q \neq 0$. Define $\mathcal{I}_{d,k}$ as the index set satisfying $\lambda_i \equiv \mu_k, i \in \mathcal{I}_{d,k}$. Further suppose that there exists an absolute constant $c_0 > 0$ such that for any d and $k \in \{0, 1, \dots, q\}$ with $\mu_k \neq 0$, we have

$$\sum_{i \in \mathcal{I}_{d,k}} \mu_k^{-s} f_i^2 \ge c_0. \tag{10}$$

Assumption 5 (a) is usually used as the traditional source condition (Caponnetto, 2006; Fischer and Steinwart, 2020, etc.). In order to obtain a reasonable lower bound, we need Assumption 5 (b). It is equivalent to assume that the $[\mathcal{H}]^s$ norm of the projection of f_{ρ}^* on the first q-th eigenspace is non-vanishing. Similar assumptions have been adopted when one interested in the lower bound of generalization error in the fixed-dimensional setting, e.g., Eq.(8) in Cui et al. (2021) and Assumption 3 in Li et al. (2023c). In a word, recalling definition 3, Assumption 5 implies that $f_{\rho}^* \in [\mathcal{H}]^s$ and $f_{\rho}^* \notin [\mathcal{H}]^t$ for any t > s.

Now we are ready to state two theorems about the exact convergence rates of the generalization error of KRR, which deal with two different ranges of source condition: $s \ge 1$ and 0 < s < 1.

Theorem 2 (Exact convergence rates when s \geq **1)** Let $c_1d^{\gamma} \leq n \leq c_2d^{\gamma}$ for some fixed $\gamma > 0$ and absolute constants c_1, c_2 . Consider $\mathcal{X} = \mathbb{S}^d$ and the marginal distribution μ to be the uniform distribution. Let $k = \{k_d\}_{d=1}^{\infty}$ be a sequence of inner product kernels on the sphere satisfying Assumption 1 and 4. Further suppose that Assumption 2 holds and Assumption 5 holds for some $s \geq 1$. Let \hat{f}_{λ} be the KRR estimator defined by (1). Define $\tilde{s} = \min\{s, 2\}$, then we have:

(i) When $\gamma \in (p + p\tilde{s}, p + p\tilde{s} + 1]$ for some $p \in \mathbb{N}$, by choosing $\lambda = d^{-\frac{\gamma + p - p\tilde{s}}{2}} \cdot \mathbf{1}_{p>0} + d^{-\frac{\gamma}{2}} \ln d \cdot \mathbf{1}_{p=0}$, we have

$$\mathbb{E}\left[\left\|\hat{f}_{\lambda} - f_{\rho}^{*}\right\|_{L^{2}}^{2} \mid \boldsymbol{X}\right] = \begin{cases} \Theta_{\mathbb{P}}\left(d^{-\gamma}\ln^{2}d\right) = \Theta_{\mathbb{P}}\left(n^{-1}\ln^{2}n\right), & p = 0, \\ \Theta_{\mathbb{P}}\left(d^{-\gamma+p}\right) = \Theta_{\mathbb{P}}\left(n^{-1+\frac{p}{\gamma}}\right), & p > 0; \end{cases}$$
(11)

(ii) When $\gamma \in (p + p\tilde{s} + 1, \ p + p\tilde{s} + 2\tilde{s} - 1]$ for some $p \in \mathbb{N}$, by choosing $\lambda = d^{-\frac{\gamma + 3p - p\tilde{s} + 1}{4}}$, we have

$$\mathbb{E}\left[\left\|\hat{f}_{\lambda} - f_{\rho}^{*}\right\|_{L^{2}}^{2} \mid \boldsymbol{X}\right] = \Theta_{\mathbb{P}}\left(d^{-\frac{\gamma - p + p\tilde{s} + 1}{2}}\right) = \Theta_{\mathbb{P}}\left(n^{-\frac{\gamma - p + p\tilde{s} + 1}{2\gamma}}\right); \tag{12}$$

(iii) When $\gamma \in (p+p\tilde{s}+2\tilde{s}-1,\ (p+1)+(p+1)\tilde{s}]$ for some $p \in \mathbb{N}$, by choosing $\lambda = d^{-\frac{\gamma+(p+1)(1-\tilde{s})}{2}}$, we have

$$\mathbb{E}\left[\left\|\hat{f}_{\lambda} - f_{\rho}^{*}\right\|_{L^{2}}^{2} \mid \boldsymbol{X}\right] = \Theta_{\mathbb{P}}\left(d^{-(p+1)\tilde{s}}\right) = \Theta_{\mathbb{P}}\left(n^{-\frac{(p+1)\tilde{s}}{\gamma}}\right). \tag{13}$$

The notation $\Theta_{\mathbb{P}}$ involves constants only depending on $s, \sigma, \gamma, c_0, \kappa, c_1$ and c_2 . In addition, the convergence rates of the generalization error of KRR can not be faster than above for any choice of regularization parameter $\lambda = \lambda(d, n) \to 0$.

Theorem 3 (Exact convergence rates when 0 < s < 1) Let $c_1 d^{\gamma} \le n \le c_2 d^{\gamma}$ for some fixed $\gamma > 0$ and absolute constants c_1, c_2 . Consider $\mathcal{X} = \mathbb{S}^d$ and the marginal distribution μ to be the uniform distribution. Let $k = \{k_d\}_{d=1}^{\infty}$ be a sequence of inner product kernels on the sphere satisfying Assumption 1 and 4. Further suppose that Assumption 2 holds and Assumption 5 holds for some 0 < s < 1. Let \hat{f}_{λ} be the KRR estimator defined by (1). Then we have:

- If $\frac{1}{2} < s < 1$:
 - (i) When $\gamma \in (p+ps, p+ps+s]$ for some $p \in \mathbb{N}$, by choosing $\lambda = d^{-\frac{\gamma+p-ps}{2}} \cdot \mathbf{1}_{p>0} + d^{-\frac{\gamma}{2}} \ln d \cdot \mathbf{1}_{p=0}$, we have

$$\mathbb{E}\left[\left\|\hat{f}_{\lambda} - f_{\rho}^{*}\right\|_{L^{2}}^{2} \mid \boldsymbol{X}\right] = \begin{cases} \Theta_{\mathbb{P}}\left(d^{-\gamma}\ln^{2}d\right) = \Theta_{\mathbb{P}}\left(n^{-1}\ln^{2}n\right), & p = 0, \\ \Theta_{\mathbb{P}}\left(d^{-\gamma+p}\right) = \Theta_{\mathbb{P}}\left(n^{-1+\frac{p}{\gamma}}\right), & p > 0; \end{cases}$$
(14)

(ii) When $\gamma \in (p+ps+s, (p+1)+(p+1)s]$ for some $p \in \mathbb{N}$, by choosing $\lambda = d^{-\frac{2p+s}{2}}$, we have $\mathbb{E}\left[\left\|\hat{f}_{\lambda} - f_{\rho}^{*}\right\|_{L^{2}}^{2} \mid \boldsymbol{X}\right] = \Theta_{\mathbb{P}}\left(d^{-(p+1)s}\right) = \Theta_{\mathbb{P}}\left(n^{-\frac{(p+1)s}{\gamma}}\right); \tag{15}$

The notation $\Theta_{\mathbb{P}}$ involves constants only depending on $s, \sigma, \gamma, c_0, \kappa, c_1$ and c_2 .

• If $0 < s \le \frac{1}{2}$: we have the same convergence rates as the case $s \in (\frac{1}{2}, 1)$ for those

$$\gamma > \frac{3s}{2(s+1)}.$$

Remark 4 For technical reasons, when $0 < s \le 1/2$, we only prove the convergence rates for those $\gamma > 3s/2(s+1)$. Note that we have 3s/2(s+1) < 1/2 when $0 < s \le 1/2$; and $3s/2(s+1) \to 0$ when $s \to 0$. Therefore, we have actually proved for almost all $\gamma > 0$.

Note that Theorem 2 and Theorem 3 show exact convergence rates (both upper and lower bounds) of KRR's generalization error, which is a much stronger result than only proving an upper bound. As we will see in Appendix B.4, since $||f_{\rho}^*||_{L^{\infty}}$ could be infinite when s < 1 thus $\mathcal{M}_1(\lambda)$ could be infinite, the proof of Theorem 3 requires a little more technique. In addition, we will prove in Theorem 5 that the rates in Theorem 3 $(s \le 1)$ achieve the minimax lower bound. Together with the statement at the end of Theorem 2, we actually prove that

the rates in Theorem 2 and Theorem 3 are the fastest convergence rates that KRR can achieve.

Next, we will state the minimax lower bound in the same large-dimensional and source condition setting as Theorem 2 and Theorem 3.

Theorem 5 (Minimax lower bound) Let $c_1d^{\gamma} \leq n \leq c_2d^{\gamma}$ for some fixed $\gamma > 0$ and absolute constants c_1, c_2 . Consider $\mathcal{X} = \mathbb{S}^d$ and the marginal distribution μ to be the uniform distribution. Let $k = \{k_d\}_{d=1}^{\infty}$ be a sequence of inner product kernels on the sphere satisfying Assumption 1 and 4. Let \mathcal{P} consist of all the distributions ρ on $\mathcal{X} \times \mathcal{Y}$ such that Assumption 2 holds and Assumption 5 holds for some s > 0. Then we have:

(i) When $\gamma \in (p + ps, p + ps + s]$ for some $p \in \mathbb{N}$, for any $\epsilon > 0$, there exist constants \mathfrak{C}_1 and \mathfrak{C} only depending on $s, \epsilon, \gamma, \sigma, \kappa, c_1$ and c_2 such that for any $d \geq \mathfrak{C}$, we have:

$$\min_{\hat{f}} \max_{\rho \in \mathcal{P}} \mathbb{E}_{(\boldsymbol{X}, \mathbf{y}) \sim \rho^{\otimes n}} \left\| \hat{f} - f_{\rho}^* \right\|_{L^2}^2 \ge \mathfrak{C}_1 d^{-\gamma + p - \epsilon}; \tag{16}$$

(ii) When $\gamma \in (p+ps+s,(p+1)+(p+1)s]$ for some $p \in \mathbb{N}$, there exist constants \mathfrak{C}_1 and \mathfrak{C} only depending on $s, \gamma, \sigma, \kappa, c_1$ and c_2 such that for any $d \geq \mathfrak{C}$, we have:

$$\min_{\hat{f}} \max_{\rho \in \mathcal{P}} \mathbb{E}_{(\boldsymbol{X}, \mathbf{y}) \sim \rho^{\otimes n}} \left\| \hat{f} - f_{\rho}^* \right\|_{L^2}^2 \ge \mathfrak{C}_1 d^{-(p+1)s}; \tag{17}$$

Theorem 5 states that there is no estimator (or learning method) that can achieve faster convergence rates than (16) and (17).

Summarizing the results in Theorem 2, Theorem 3 and Theorem 5, Figure 2 shows the convergence rates of KRR and corresponding minimax lower rates with respect to dimension d for any $\gamma > 0$. We can see that the rates decrease when the scaling γ increases, indicating that the performance becomes better when the sample size n grows. Moreover, we can observe several intriguing phenomena.

Curve's evolution with source condition. Since we consider source condition s > 0, we can compare the rate curves in Figure 2 for different s and see how they evolve with s.

Let us first see the minimax lower rates. For any s > 0, there are 2 periods with respect to the value of γ : The length of the first period, i.e., $(p+ps, p+ps+s], p \in \mathbb{N}$, will decrease to 0 as s getting close to 0; The length of the second period, i.e., (p+ps+s, (p+1)+(p+1)s], equals 1 for all s > 0.

Next, we see the convergence rates of KRR, which is more intriguing.

- When $0 < s \le 1$, there are 2 periods with respect to the value of γ and the curve is the same as the minimax lower rates. (In fact, Theorem 3 only proves the results for $\gamma > 3s/2(s+1)$ when $s \le 1/2$, we write $\gamma > 0$ with a little bit of notation abusement.)
- When 1 < s < 2, there are 3 periods with respect to the value of γ : The length of the first period, i.e., (p+ps,p+ps+1], equals 1 for all 1 < s < 2; The length of the second period, i.e., (p+ps+1,p+ps+2s-1] is 2s-2, thus this period will degenerate as s getting close to 1; The length of the third period, i.e., (p+ps+2s-1,(p+1)+(p+1)s] is 2-s, thus this period will degenerate as s getting close to 2.

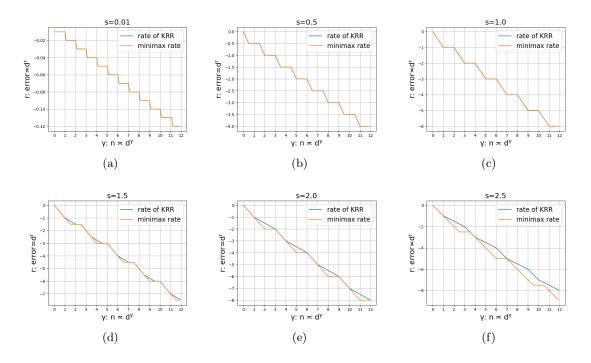


Figure 2: Convergence rates of KRR in Theorem 2, Theorem 3 and corresponding minimax lower rates in Theorem 5 (ignoring a ϵ -difference) with respect to dimension d. We present 6 graphs corresponding to 6 kinds of source conditions: s = 0.01, 0.5, 1.0, 1.5, 2.0, 2.5. The x-axis represents asymptotic scaling, $\gamma : n \times d^{\gamma}$; the y-axis represents the convergence rate of generalization error, $r : \text{error} \times d^r$.

• When $s \ge 2$, the curve does not change with s and there are 2 periods with respect to the value of γ : The length of the first period, i.e., (3p, 3p + 1], equals 1 for all $s \ge 2$; The length of the second period, i.e., (3p + 1, 3p + 3], equals 2 for all $s \ge 2$.

Minimax optimality and new saturation effect of KRR. As can be seen in Figure 2 (a)(b)(c), the convergence rates of KRR match the minimax lower bound for all $\gamma > 0$, thus we prove the minimax optimality of KRR when $0 < s \le 1$. In contrast, when s > 1, Figure 2 (d)(e)(f) illustrate that KRR can not achieve the minimax lower bound in Theorem 5 for certain ranges of γ , which we refer to as the new saturation effect of KRR. We will discuss the implications of this new saturation effect in the following.

In the fixed-dimensional setting, kernel ridge regression has been studied as a special kind of spectral algorithm (Gerfo et al., 2008). Each spectral algorithm is determined by a filter function and the difference between spectral algorithms is characterized by an index called the "qualification" (τ) of the filter function (see, e.g., Zhang et al. 2023a, Definition 1). Since KRR has qualification $\tau = 1$ and gradient flow has qualification $\tau = \infty$ (Zhang et al., 2023a, Example 1, 2), the main difference between KRR and gradient flow is the saturation effect (Li et al., 2023b). It says when $s > 2\tau = 2$, no matter how carefully one

tunes the KRR, the convergence rate can not be faster than $n^{-\frac{2\beta}{2\beta+1}}$, thus can not achieve the minimax lower bound $n^{-\frac{s\beta}{s\beta+1}}$.

In the large-dimensional setting and for inner product kernel on the sphere, our results show that the saturation effect of KRR happens in a new regime $1 < s \le 2 = 2\tau$. In addition, we conjecture that there are other spectral algorithms (e.g., gradient flow) that can achieve the minimax lower bound in Theorem 5 for all s > 0. This new saturation effect strongly suggests that qualification (τ) itself is insufficient to characterize the filter function (or spectral algorithm) in the large-dimensional setting.

Periodic plateau behavior. If 0 < s < 2, Figure 2 (a)(b)(c) show that when γ varies within certain ranges, the value of vertical axis, r, does not change. We refer to such ranges of γ as the plateau period. When s exceeds 2, the plateau period of KRR's convergence rates degenerates and the plateau period of minimax lower rates still exists. Also note that the length of each plateau period varies with the values s > 0.

For these plateau periods, if we fix a large dimension d and increase γ (or equivalently, increase the sample size n), the convergence rates of KRR or minimax lower rates stay invariant in certain ranges. Therefore, in order to improve the rate, one has to increase the sample size above a certain threshold.

Figure 3 provides an alternative representation of our results, which shows the convergence rates of KRR and corresponding minimax lower rates with respect to sample size n. We can observe the "multiple descent behavior" (for both the convergence rates of KRR and the minimax lower rates) from Figure 3.

Multiple descent behavior. Let us first see the minimax lower rates. For any s > 0, the curve achieves its peaks at $\gamma = p + ps, p \in \mathbb{N}^+$, and achieve its isolated valleys at $\gamma = p + ps + s, p \in \mathbb{N}^+$.

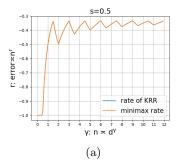
For the convergence rates of KRR:

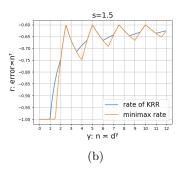
- When $0 < s \le 1$, the curve is the same as the curve of minimax lower rates.
- When 1 < s < 2, the curve achieves its peaks at $\gamma = p + p\tilde{s}, p \in \mathbb{N}^+$; achieve its isolated valleys at $\gamma = p + p\tilde{s} + 1, p \in \mathbb{N}^+$ and achieve its hillside at $\gamma = p + p\tilde{s} + 2\tilde{s} 1, p \in \mathbb{N}^+$.
- When $s \ge 2$, the curve does not change with s, which achieves its peaks at $\gamma = 3p, p \in \mathbb{N}^+$, and achieve its isolated valleys at $\gamma = 3p + 1, p \in \mathbb{N}^+$.

3.3 Applications to neural tangent kernel

In this subsection, we look at a specific example, i.e., the neural tangent kernel (NTK) of a two-layer fully connected ReLU neural network $k_d = k_d^{\text{NT}}$. Still, we suppose that $\mathcal{X} = \mathbb{S}^d$ and μ is uniform distribution on \mathbb{S}^d . It has been shown in Bietti and Mairal (2019); Lu et al. (2023) that NTK is an example of inner product kernel satisfying

$$k_d^{\text{NT}}(\boldsymbol{x}, \boldsymbol{x}') = \Phi\left(\langle \boldsymbol{x}, \boldsymbol{x}' \rangle\right), \ \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{S}^d,$$





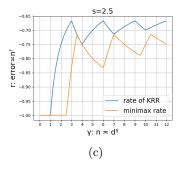


Figure 3: Convergence rates of KRR in Theorem 2, Theorem 3 and corresponding minimax lower rates in Theorem 5 (ignoring a ϵ -difference) with respect to sample size n. We present 3 graphs corresponding to 3 kinds of source conditions: s = 0.5, 1.5, 2.5. The x-axis represents asymptotic scaling, $\gamma : n \approx d^{\gamma}$; the y-axis represents the convergence rate of generalization error, $r : \text{error} \approx n^r$.

where $\Phi(t) = \sum_{j=0}^{\infty} a_j t^j \in \mathcal{C}^{\infty}$ ([-1,1]) is a fixed function independent of d and $a_0 > 0$; $a_1 > 0$; $a_j > 0$, $\forall j = 2, 4, 6 \cdots$; $a_j = 0$, $\forall j = 3, 5, 7, \cdots$

Lemma 5 in Lu et al. (2023) also showed that $\sup_{\boldsymbol{x} \in \mathcal{X}} k_d^{\text{NT}}(\boldsymbol{x}, \boldsymbol{x}) \leq 1$.

For the neural tangent kernel $k_d^{\rm NT}$, the following theorems provide the exact convergence rates of the generalization error of KRR and the corresponding minimax lower bound. Since the proofs are similar to Theorem 2, Theorem 3 and Theorem 5, we omit the proofs of the following theorems.

Theorem 6 (NTK: exact convergence rates when $s \ge 1$) Let $c_1d^{\gamma} \le n \le c_2d^{\gamma}$ for some fixed $\gamma > 0$ and absolute constants c_1, c_2 . Consider $\mathcal{X} = \mathbb{S}^d$ and the marginal distribution μ to be the uniform distribution. Let $k = \{k_d^{\mathrm{NT}}\}_{d=1}^{\infty}$ be a sequence of neural tangent kernels of a two-layer fully connected ReLU neural network on the sphere. Further suppose that Assumption 2 holds and Assumption 5 holds for some $s \ge 1$. Let \hat{f}_{λ} be the KRR estimator defined by (1). For any $p \in \mathcal{I}_p$, we define p' = p + 2, if $p \ge 2$; and define p' = p + 1, if $p \le 1$. Define $\tilde{s} = \min\{s, 2\}$, then we have:

(i) When $\gamma \in (p + p\tilde{s}, p' + p\tilde{s}]$ for some $p \in \mathcal{I}_p$, by choosing $\lambda = d^{-\frac{\gamma + p - p\tilde{s}}{2}} \cdot \mathbf{1}_{p>0} + d^{-\frac{\gamma}{2}} \ln d \cdot \mathbf{1}_{p=0}$, we have

$$\mathbb{E}\left[\left\|\hat{f}_{\lambda} - f_{\rho}^{*}\right\|_{L^{2}}^{2} \mid \boldsymbol{X}\right] = \begin{cases} \Theta_{\mathbb{P}}\left(d^{-\gamma}\ln^{2}d\right) = \Theta_{\mathbb{P}}\left(n^{-1}\ln^{2}n\right), & p = 0, \\ \Theta_{\mathbb{P}}\left(d^{-\gamma+p}\right) = \Theta_{\mathbb{P}}\left(n^{-1+\frac{p}{\gamma}}\right), & p > 0; \end{cases}$$
(18)

(ii) When $\gamma \in (p' + p\tilde{s}, 2p'\tilde{s} - p' + 2p - p\tilde{s}]$ for some $p \in \mathcal{I}_p$, by choosing $\lambda = d^{-\frac{\gamma + p' + 2p - p\tilde{s}}{4}}$, we have

$$\mathbb{E}\left[\left\|\hat{f}_{\lambda} - f_{\rho}^{*}\right\|_{L^{2}}^{2} \mid \boldsymbol{X}\right] = \Theta_{\mathbb{P}}\left(d^{-\frac{\gamma}{2} + \frac{p'}{2} - \frac{p\tilde{s}}{2} - 2}\right) = \Theta_{\mathbb{P}}\left(n^{-\frac{1}{2} + \frac{p'}{2\gamma} - \frac{p\tilde{s}}{2\gamma} - \frac{2}{\gamma}}\right); \tag{19}$$

(iii) When $\gamma \in (2p'\tilde{s} - p' + 2p - p\tilde{s}, p' + p'\tilde{s}]$ for some $p \in \mathcal{I}_p$, by choosing $\lambda = d^{-\frac{\gamma + p'(1 - \tilde{s})}{2}}$, we have

$$\mathbb{E}\left[\left\|\hat{f}_{\lambda} - f_{\rho}^{*}\right\|_{L^{2}}^{2} \mid \boldsymbol{X}\right] = \Theta_{\mathbb{P}}\left(d^{-p'\tilde{s}}\right) = \Theta_{\mathbb{P}}\left(n^{-\frac{p'\tilde{s}}{\gamma}}\right). \tag{20}$$

The notation $\Theta_{\mathbb{P}}$ involves constants only depending on $s, \sigma, \gamma, c_0, c_1$ and c_2 .

Theorem 7 (NTK: exact convergence rates when 0 < s < 1) Let $c_1 d^{\gamma} \le n \le c_2 d^{\gamma}$ for some fixed $\gamma > 0$ and absolute constants c_1, c_2 . Consider $\mathcal{X} = \mathbb{S}^d$ and the marginal distribution μ to be the uniform distribution. Let $k = \{k_d^{\text{NT}}\}_{d=1}^{\infty}$ be a sequence of neural tangent kernels of a two-layer fully connected ReLU neural network on the sphere. Further suppose that Assumption 2 holds and Assumption 5 holds for some 0 < s < 1. Let \hat{f}_{λ} be the KRR estimator defined by (1). Denote $\mathcal{I}_p = \{0,1\} \cup \{2,4,6,\cdots\}$. For any $p \in \mathcal{I}_p$, we define p' = p + 2, if $p \ge 2$; and define p' = p + 1, if $p \le 1$. Then we have:

- If $\frac{1}{2} < s < 1$:
 - (i) When $\gamma \in (p+ps, p+p's]$ for some $p \in \mathcal{I}_p$, by choosing $\lambda = d^{-\frac{\gamma+p-ps}{2}} \cdot \mathbf{1}_{p>0} + d^{-\frac{\gamma}{2}} \ln d \cdot \mathbf{1}_{p=0}$, we have

$$\mathbb{E}\left[\left\|\hat{f}_{\lambda} - f_{\rho}^{*}\right\|_{L^{2}}^{2} \mid \boldsymbol{X}\right] = \begin{cases} \Theta_{\mathbb{P}}\left(d^{-\gamma}\ln^{2}d\right) = \Theta_{\mathbb{P}}\left(n^{-1}\ln^{2}n\right), & p = 0, \\ \Theta_{\mathbb{P}}\left(d^{-\gamma+p}\right) = \Theta_{\mathbb{P}}\left(n^{-1+\frac{p}{\gamma}}\right), & p > 0; \end{cases}$$
(21)

(ii) When $\gamma \in (p + p's, p' + p's]$ for some $p \in \mathcal{I}_p$, by choosing $\lambda = d^{-p - \frac{(p'-p)s}{2}}$, we have

$$\mathbb{E}\left[\left\|\hat{f}_{\lambda} - f_{\rho}^{*}\right\|_{L^{2}}^{2} \mid \boldsymbol{X}\right] = \Theta_{\mathbb{P}}\left(d^{-p's}\right) = \Theta_{\mathbb{P}}\left(n^{-\frac{p's}{\gamma}}\right); \tag{22}$$

The notation $\Theta_{\mathbb{P}}$ involves constants only depending on $s, \sigma, \gamma, c_0, c_1$ and c_2 .

• If $0 < s \le \frac{1}{2}$: we have the same convergence rates as the case $s \in (\frac{1}{2}, 1)$ for those

$$\gamma > \frac{3s}{2(s+1)}.$$

Theorem 8 (NTK: minimax lower bound) Let $c_1d^{\gamma} \leq n \leq c_2d^{\gamma}$ for some fixed $\gamma > 0$ and absolute constants c_1, c_2 . Consider $\mathcal{X} = \mathbb{S}^d$ and the marginal distribution μ to be the uniform distribution. Let $k = \{k_d^{\text{NT}}\}_{d=1}^{\infty}$ be a sequence of neural tangent kernels of a two-layer fully connected ReLU neural network on the sphere. Let \mathcal{P} consist of all the distributions ρ on $\mathcal{X} \times \mathcal{Y}$ such that Assumption 2 holds and Assumption 5 holds for some s > 0. Then we have:

(i) When $\gamma \in (p + ps, p + p's]$ for some $p \in \mathcal{I}_p$, for any $\epsilon > 0$, there exist constants \mathfrak{C}_1 and \mathfrak{C} only depending on $s, \epsilon, \gamma, \sigma, c_1$ and c_2 such that for any $d \geq \mathfrak{C}$, we have:

$$\min_{\hat{f}} \max_{\rho \in \mathcal{P}} \mathbb{E}_{(\boldsymbol{X}, \mathbf{y}) \sim \rho^{\otimes n}} \left\| \hat{f} - f_{\rho}^* \right\|_{L^2}^2 \ge \mathfrak{C}_1 d^{-\gamma + p - \epsilon}; \tag{23}$$

(ii) When $\gamma \in (p+p's, p'+p's]$ for some $p \in \mathcal{I}_p$, there exist constants \mathfrak{C}_1 and \mathfrak{C} only depending on s, γ, σ, c_1 and c_2 such that for any $d \geq \mathfrak{C}$, we have:

$$\min_{\hat{f}} \max_{\rho \in \mathcal{P}} \mathbb{E}_{(\boldsymbol{X}, \mathbf{y}) \sim \rho^{\otimes n}} \left\| \hat{f} - f_{\rho}^* \right\|_{L^2}^2 \ge \mathfrak{C}_1 d^{-p's}; \tag{24}$$

Note that the results in this subsection will be the same as the theorems in Section 3.2 if we change the above definition of p' to $p' = p + 1, \forall p = 0, 1, 2, \cdots$.

4. Conclusion and discussion

In this paper, we first establish a new framework for studying the asymptotic generalization error of kernel ridge regression (Theorem 1). This framework makes few assumptions on the RKHS, the true function and the relation between d and n, thus it is suitable for studying various topics about KRR's generalization error in both fixed-dimensional and large-dimensional settings. Moreover, the results in Theorem 1 provides the matching upper and lower bounds of the generalization error with certain regularization parameter, which is more instructive than just the upper bound.

Based on this framework, we then consider inner product kernel on the sphere and the large-dimensional setting $(n \approx d^{\gamma}, \gamma > 0)$. Given that f_{ρ}^* falls into $[\mathcal{H}]^s$, an interpolation space of RKHS, Theorem 2 and Theorem 3 prove the exact convergence rates of KRR's generalization error under the best choice of regularization parameter and Theorem 5 proves the corresponding minimax lower bound. These results show the minimax optimality of KRR when $0 < s \le 1$ and the new saturation effect of KRR s > 1. We also discuss how the rate curves (varying along γ) evolve with the value of s and discuss the "periodic plateau behavior" and "multiple descent behavior" of KRR in the large-dimensional setting.

Similar periodic behavior has been observed for kernel methods in the large-dimensional setting in related literature, we next make some discussion on these works. There is a line of work studying the inconsistency of kernel methods with inner product kernels in the large-dimensional setting $n \approx d^{\gamma}$, $\gamma > 0$ (Ghorbani et al., 2021; Mei et al., 2022; Misiakiewicz, 2022, etc.). Assuming the true function f_{ρ}^{*} to be square-integrable (or equivalently s = 0 in our setting) on the sphere, Ghorbani et al. (2021, Theorem 4) proves that the generalization error of KRR $R_{\rm KR}$ (f_{ρ}^{*} , X, λ) satisfies (with high probability)

$$\left| R_{KR} \left(f_{\rho}^*, \boldsymbol{X}, \lambda \right) - \left\| P_{>\ell} f_{\rho}^* \right\|_{L^2}^2 \right| \le \epsilon \left(\left\| f_{\rho}^* \right\|_{L^2}^2 + \sigma^2 \right), \forall \ 0 < \lambda < \lambda^*, \tag{25}$$

where $\ell = \lfloor \gamma \rfloor$ is the greatest integer that is less or equal to γ , $P_{>\ell}$ means the projection onto polynomials with degree $> \ell$, ϵ is any positive real number and λ^* is defined as Ghorbani et al. 2021, Eq.(20). (25) implies that generalization error will drop when γ exceeds an integer and stay invariant for other γ (see the cartoon representation in Ghorbani et al. 2021, Figure 5). In our paper, when s > 0 and efficiently close to 0, similar behavior has been observed in Figure 2 (a) that the convergence rate drops abruptly around each integer $\gamma \in \mathbb{N}$.

A more recent work Lu et al. (2023) considers the optimality of early stopping kernel gradient flow in the same large-dimensional setting $c_1d^{\gamma} \leq n \leq c_2d^{\gamma}$, $\gamma > 0$. They also consider inner product kernel on the sphere and assume that the true function falls into

the RKHS $f_{\rho}^* \in \mathcal{H}$ (or equivalently, s = 1). Denoting $p = \lfloor \gamma/2 \rfloor$, Lu et al. 2023, Theorem 4.3 proves that by properly choosing the early stopping time \widehat{T} , the upper bound of the convergence rate is:

• When $\gamma \in \{2, 4, 6, \dots\}$, then, there exist constants \mathfrak{C} and \mathfrak{C}_i , where i = 1, 2, 3, only depending on γ , c_1 , and c_2 , such that for any $d \geq \mathfrak{C}$, we have

$$\|f_{\widehat{T}} - f_{\star}\|_{L^{2}}^{2} \le \mathfrak{C}_{1} n^{-\frac{1}{2}} \tag{26}$$

holds with probability at least $1 - \mathfrak{C}_2 \exp\{-\mathfrak{C}_3 n^{1/2}\}$.

• When $\gamma \in \bigcup_{j=0}^{\infty} (2j, 2j+1]$, for any $\delta > 0$, there exist constants \mathfrak{C} and \mathfrak{C}_i , where i=1,2,3, only depending on γ , δ , c_1 , and c_2 , such that for any $d \geq \mathfrak{C}$, we have

$$\left\| f_{\widehat{T}} - f_{\rho}^* \right\|_{L^2}^2 \le \mathfrak{C}_1 n^{-\frac{\gamma - p}{\gamma}} \log(n) \tag{27}$$

holds with probability at least $1 - \delta - \mathfrak{C}_2 \exp\{-\mathfrak{C}_3 n^{p/\gamma} \log(n)\}$.

• When $\gamma \in \bigcup_{j=0}^{\infty} (2j+1,2j+2)$, then, for any $\delta > 0$, there exist constants \mathfrak{C} and \mathfrak{C}_i , where i = 1, 2, 3, only depending on γ , δ , c_1 , and c_2 , such that for any $d \geq \mathfrak{C}$, we have

$$\left\| f_{\widehat{T}} - f_{\star} \right\|_{L^{2}}^{2} \le \mathfrak{C}_{1} n^{-\frac{p+1}{\gamma}} \tag{28}$$

holds with probability at least $1 - \delta - \mathfrak{C}_2 \exp\{-\mathfrak{C}_3 n^{1-(p+1)/\gamma}\}$.

Ignoring the log term, simple calculation shows that the convergence rate is consistent with the rate in Theorem 2 when s=1 (see, e.g., Figure 2 (c)). Lu et al. (2023) also proves that the above upper bound matches the minimax lower bound, thus proving the minimax optimality of early stopping kernel gradient flow under the assumption $f_{\rho}^* \in \mathcal{H}$. In contrast to Lu et al. (2023), we provide not only the upper bound but also the lower bound of the convergence rates of KRR under general source condition s>0. We have seen that the "periodic plateau behavior" and "multiple descent behavior" observed in Lu et al. (2023) still exist for s>0 and the plateau length will change with the value of s. By considering source condition s>0, our restriction on the true function is much milder, thus we provide a more complete characterization of the generalization error of KRR.

The periodic behavior in large dimension has also been observed for "kernel interpolation estimator", for instance, Liang et al. (2020) for the inner product kernel and Aerni et al. (2022) for the convolutional kernel. Although technically complicated, a direct follow-up question is the convergence rate of generalization error for general kernels and domains. We believe that it is an interesting research direction to study the generalization behavior of kernel methods in the large-dimensional setting, which will exhibit a wealth of new phenomena compared with the fixed-dimensional setting.

Acknowledgments and Disclosure of Funding

Qian Lin is supported in part by National Natural Science Foundation of China (Grant 92370122, Grant 11971257) and the Beijing Natural Science Foundation (Grant Z190001).

Appendix

In the Appendix, we provide the proof of Theorem 1 (Appendix A), Theorem 2 & 3 (Appendix B) and Theorem 5 (Appendix C). Necessary auxiliary results can be found in Appendix D.

A. Proof of Theorem 1

The proof of Theorem 1 consists of the following steps: First, we introduce the bias-variance decomposition in Section A.1. Next, we derive the bounds of variance term in Section A.2 and bias term in Section A.3. Finally, using the results in these sections, we formally prove Theorem 1 in Section A.4.

A.1 Bias-variance decomposition

The proof of Theorem 1 is based on the traditional bias-variance decomposition. The contribution in this paper is that we refine the tools in Li et al. (2023a) and Li et al. (2023c) to handle the large-dimensional case. Throughout the proof, we denote

$$T_{\lambda} = T + \lambda; \quad T_{X\lambda} = T_X + \lambda,$$
 (29)

where λ is the regularization parameter. We use $\|\cdot\|_{\mathscr{B}(B_1,B_2)}$ to denote the operator norm of a bounded linear operator from a Banach space B_1 to B_2 , i.e., $\|A\|_{\mathscr{B}(B_1,B_2)} = \sup_{\|f\|_{B_1}=1} \|Af\|_{B_2}$.

Without bringing ambiguity, we will briefly denote the operator norm as $\|\cdot\|$. In addition, we use $\operatorname{tr} A$ and $\|A\|_1$ to denote the trace and the trace norm of an operator. We use $\|A\|_2$ to denote the Hilbert-Schmidt norm. In addition, we denote $L^2(\mathcal{X}, \mu)$ as L^2 , $L^{\infty}(\mathcal{X}, \mu)$ as L^{∞} for brevity throughout the proof.

We also need the following essential notations in our proof, which are frequently used in related literature. Denote the samples $\mathbf{Z} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$. Define the sampling operator $K_{\mathbf{x}} : \mathbb{R} \to \mathcal{H}, \ y \mapsto yk(\mathbf{x}, \cdot)$ and its adjoint operator $K_{\mathbf{x}}^* : \mathcal{H} \to \mathbb{R}, \ f \mapsto f(\mathbf{x})$. Then we can define $T_{\mathbf{x}} = K_{\mathbf{x}}K_{\mathbf{x}}^*$. Further, we define the sample covariance operator $T_{\mathbf{X}} : \mathcal{H} \to \mathcal{H}$ as

$$T_{\mathbf{X}} := \frac{1}{n} \sum_{i=1}^{n} K_{\mathbf{x}_i} K_{\mathbf{x}_i}^*. \tag{30}$$

Then we know that $||T_{\boldsymbol{X}}|| \le ||T_{\boldsymbol{X}}||_1 \le \kappa^2$ and $T_{\boldsymbol{X}}$ is a trace class thus compact operator. Further, define the sample basis function

$$g_{\mathbf{Z}} := \frac{1}{n} \sum_{i=1}^{n} K_{\mathbf{x}_i} y_i \in \mathcal{H}. \tag{31}$$

As shown in Caponnetto and de Vito (2007), the operator form of the KRR estimator (1) writes

$$\hat{f}_{\lambda} = (T_{\mathbf{X}} + \lambda)^{-1} g_{\mathbf{Z}}.\tag{32}$$

In order to derive the bias term, we define

$$\tilde{g}_{\mathbf{Z}} := \mathbb{E}\left(g_{\mathbf{Z}}|\mathbf{X}\right) = \frac{1}{n} \sum_{i=1}^{n} K_{\mathbf{x}_{i}} f_{\rho}^{*}(\mathbf{x}_{i}) \in \mathcal{H}; \tag{33}$$

and

$$\tilde{f}_{\lambda} := \mathbb{E}\left(\hat{f}_{\lambda}|\mathbf{X}\right) = (T_{\mathbf{X}} + \lambda)^{-1}\,\tilde{g}_{\mathbf{Z}} \in \mathcal{H}.$$
 (34)

We also need to define the expectation of $g_{\mathbf{Z}}$ as

$$g = \mathbb{E}g_{\mathbf{Z}} = \int_{\mathcal{X}} k(\mathbf{x}, \cdot) f_{\rho}^{*}(\mathbf{x}) d\mu(\mathbf{x}) = S_{k}^{*} f_{\rho}^{*} \in \mathcal{H},$$
(35)

and

$$f_{\lambda} = (T + \lambda)^{-1} g = (T + \lambda)^{-1} S_k^* f_{\rho}^*.$$
 (36)

Then we have the decomposition

$$\hat{f}_{\lambda} - f_{\rho}^{*} = \frac{1}{n} (T_{\mathbf{X}} + \lambda)^{-1} \sum_{i=1}^{n} K_{\mathbf{x}_{i}} y_{i} - f_{\rho}^{*}$$

$$= \frac{1}{n} (T_{\mathbf{X}} + \lambda)^{-1} \sum_{i=1}^{n} K_{\mathbf{x}_{i}} (f_{\rho}^{*}(\mathbf{x}_{i}) + \epsilon_{i}) - f_{\rho}^{*}$$

$$= (T_{\mathbf{X}} + \lambda)^{-1} \tilde{g}_{\mathbf{Z}} + \frac{1}{n} \sum_{i=1}^{n} (T_{\mathbf{X}} + \lambda)^{-1} K_{\mathbf{x}_{i}} \epsilon_{i} - f_{\rho}^{*}$$

$$= (\tilde{f}_{\lambda} - f_{\rho}^{*}) + \frac{1}{n} \sum_{i=1}^{n} (T_{\mathbf{X}} + \lambda)^{-1} K_{\mathbf{x}_{i}} \epsilon_{i}.$$
(37)

Taking expectation over the noise ϵ conditioned on X and noticing that $\epsilon | \mathbf{x}$ are independent noise with mean 0 and variance σ^2 , we obtain the bias-variance decomposition:

$$\mathbb{E}\left(\left\|\hat{f}_{\lambda} - f_{\rho}^{*}\right\|_{L^{2}}^{2} \mid \boldsymbol{X}\right) = \mathbf{Bias}^{2}(\lambda) + \mathbf{Var}(\lambda), \tag{38}$$

where

$$\mathbf{Bias}^{2}(\lambda) := \left\| \tilde{f}_{\lambda} - f_{\rho}^{*} \right\|_{L^{2}}^{2}, \quad \mathbf{Var}(\lambda) := \frac{\sigma^{2}}{n^{2}} \sum_{i=1}^{n} \left\| (T_{\boldsymbol{X}} + \lambda)^{-1} k(\boldsymbol{x}_{i}, \cdot) \right\|_{L^{2}}^{2}. \tag{39}$$

Given the decomposition (38), we next derive the upper and lower bounds of $\mathbf{Bias}^2(\lambda)$ and $\mathbf{Var}(\lambda)$ in the following two subsections.

A.2 Variance term

In this subsection, our goal is to derive Theorem 13, which shows the upper and lower bounds of variance under some approximation conditions. Before formally introduce Theorem 13, we have a lot of preparatory work.

Following Li et al. (2023a), we consider the sample subspace

$$\mathcal{H}_n = \operatorname{span} \{k(\boldsymbol{x}_1,\cdot),\ldots,k(\boldsymbol{x}_n,\cdot)\} \subset \mathcal{H}.$$

Recall the notation $\mathbb{K}(\boldsymbol{X}, \boldsymbol{X}) = (k(\boldsymbol{x}_i, \boldsymbol{x}_j))_{n \times n}$ and $\mathbb{K}(\boldsymbol{X}, \cdot) = \{k(\boldsymbol{x}_1, \cdot), \dots, k(\boldsymbol{x}_n, \cdot)\}$. Define the normalized sample kernel matrix

$$\boldsymbol{K} = \frac{1}{n}\mathbb{K}(\boldsymbol{X}, \boldsymbol{X}).$$

Then, it is easy to verify that $\operatorname{Ran}(T_{\boldsymbol{X}}) = \mathcal{H}_n$ and \boldsymbol{K} is the representation matrix of $T_{\boldsymbol{X}}$ under the natural basis $\{k(\boldsymbol{x}_1,\cdot),\ldots,k(\boldsymbol{x}_n,\cdot)\}$. Consequently, for any continuous function φ we have

$$\varphi(T_{\mathbf{X}})\mathbb{K}(\mathbf{X},\cdot) = \varphi(\mathbf{K})\mathbb{K}(X,\cdot),\tag{40}$$

where the left-hand side means applying the operator elementwise. Since the property of reproducing kernel Hilbert space implies $\langle k(\boldsymbol{x},\cdot), f \rangle_{\mathcal{H}} = f(\boldsymbol{x}), \ \forall f \in \mathcal{H}$, taking inner product elementwise between (40) and f, we have

$$(\varphi(T_{\mathbf{X}})f)[\mathbf{X}] = \varphi(\mathbf{K})f[\mathbf{X}]. \tag{41}$$

Moreover, for $f, g \in \mathcal{H}$, we define empirical semi-inner product

$$\langle f, g \rangle_{L^2, n} = \frac{1}{n} \sum_{i=1}^n f(\boldsymbol{x}_i) g(\boldsymbol{x}_i) = \frac{1}{n} f[\boldsymbol{X}]^{\top} g[\boldsymbol{X}], \tag{42}$$

and denote by $\|\cdot\|_{L^2,n}$ the corresponding empirical semi-norm. We also denote by P_n the empirical measure with respect to $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$.

For simplicity of notations, we denote $k_{\boldsymbol{x}}(\cdot) = k(\boldsymbol{x}, \cdot), \ \boldsymbol{x} \in \mathcal{X}$ in the rest of the proof. The following lemma rewrites the variance term (39) using the empirical semi-norm.

Lemma 9 The variance term in (39) satisfies that

$$\mathbf{Var}(\lambda) = \frac{\sigma^2}{n} \int_{\mathcal{X}} \left\| (T_{\mathbf{X}} + \lambda)^{-1} k_{\mathbf{x}}(\cdot) \right\|_{L^2, n}^2 d\mu(\mathbf{x}). \tag{43}$$

Proof First, we have

$$\mathbf{Var}(\lambda) := \frac{\sigma^2}{n^2} \sum_{i=1}^n \left\| (T_{\boldsymbol{X}} + \lambda)^{-1} k(\boldsymbol{x}_i, \cdot) \right\|_{L^2}^2$$

$$= \frac{\sigma^2}{n^2} \left\| (T_{\boldsymbol{X}} + \lambda)^{-1} \mathbb{K}(\boldsymbol{X}, \cdot) \right\|_{L^2(\mathbb{R}^n)}^2$$

$$\stackrel{(40)}{=} \frac{\sigma^2}{n^2} \left\| (\boldsymbol{K} + \lambda)^{-1} \mathbb{K}(\boldsymbol{X}, \cdot) \right\|_{L^2(\mathbb{R}^n)}^2$$

$$= \frac{\sigma^2}{n^2} \int_{\mathcal{X}} \mathbb{K}(\boldsymbol{x}, \boldsymbol{X}) (\boldsymbol{K} + \lambda)^{-2} \mathbb{K}(\boldsymbol{X}, \boldsymbol{x}) \, d\mu(\boldsymbol{x}). \tag{44}$$

Next, using (41) and the fact that $k_x[X] = \mathbb{K}(X, x)$, we have

$$((T_{\boldsymbol{X}} + \lambda)^{-1}k_{\boldsymbol{x}})[\boldsymbol{X}] = (\boldsymbol{K} + \lambda)^{-1}k_{\boldsymbol{x}}[\boldsymbol{X}] = (\boldsymbol{K} + \lambda)^{-1}\mathbb{K}(\boldsymbol{X}, \boldsymbol{x}),$$

which implies

$$\frac{1}{n}\mathbb{K}(\boldsymbol{x},\boldsymbol{X})(\boldsymbol{K}+\lambda)^{-2}\mathbb{K}(\boldsymbol{X},\boldsymbol{x}) = \frac{1}{n}\left\|(\boldsymbol{K}+\lambda)^{-1}\mathbb{K}(\boldsymbol{X},\boldsymbol{x})\right\|_{\mathbb{R}^{n}}^{2}$$

$$= \frac{1}{n}\left\|\left((T_{\boldsymbol{X}}+\lambda)^{-1}k_{\boldsymbol{x}}\right)[\boldsymbol{X}]\right\|_{\mathbb{R}^{n}}^{2}$$

$$= \left\|(T_{\boldsymbol{X}}+\lambda)^{-1}k_{\boldsymbol{x}}\right\|_{L^{2}n}^{2}.$$
(45)

Therefore, plugging (45) into (44), we get the desired results

$$\mathbf{Var}(\lambda) = \frac{\sigma^2}{n} \int_{\mathcal{X}} \left\| (T_{\boldsymbol{X}} + \lambda)^{-1} k_{\boldsymbol{x}}(\cdot) \right\|_{L^2, n}^2 d\mu(\boldsymbol{x}).$$

The operator form (43) allows us to apply concentration inequalities and establish the following two-step approximation (recall the notations $T_{X\lambda}$ and T_{λ} in (29)).

$$\int_{\mathcal{X}} \left\| T_{\boldsymbol{X}\lambda}^{-1} k_{\boldsymbol{x}} \right\|_{L^{2},n}^{2} d\mu(\boldsymbol{x}) \stackrel{A}{\approx} \int_{\mathcal{X}} \left\| T_{\lambda}^{-1} k_{\boldsymbol{x}} \right\|_{L^{2},n}^{2} d\mu(\boldsymbol{x}) \stackrel{B}{\approx} \int_{\mathcal{X}} \left\| T_{\lambda}^{-1} k_{\boldsymbol{x}} \right\|_{L^{2}}^{2} d\mu(\boldsymbol{x}). \tag{46}$$

Note that the above two-step approximation is an enhanced version of approximation (S24) in Li et al. (2023a).

Approximation B. The following lemma characterizes the magnitude of Approximation B in high probability. Recall the definitions of $\mathcal{N}_1(\lambda)$ and $\mathcal{N}_2(\lambda)$ in (4).

Lemma 10 (Approximation B) Suppose that Assumption 1, 2 and 3 hold. For any $\lambda = \lambda(d, n) \to 0$ and any fixed $\delta \in (0, 1)$, when n is sufficiently large, with probability at least $1 - \delta$, we have

$$\frac{1}{2} \int_{\mathcal{X}} \|T_{\lambda}^{-1} k_{\boldsymbol{x}}\|_{L^{2}}^{2} d\mu(\boldsymbol{x}) - R_{2} \leq \int_{\mathcal{X}} \|T_{\lambda}^{-1} k_{\boldsymbol{x}}\|_{L^{2}, n}^{2} d\mu(\boldsymbol{x}) \leq \frac{3}{2} \int_{\mathcal{X}} \|T_{\lambda}^{-1} k_{\boldsymbol{x}}\|_{L^{2}}^{2} d\mu(\boldsymbol{x}) + R_{2},$$
(47)

where

$$R_2 = \frac{5\mathcal{N}_2(\lambda)}{3n} \ln \frac{2}{\delta}.$$
 (48)

Proof Define a function

$$f(z) = \int_{\mathcal{X}} \left(T_{\lambda}^{-1} k_{x}(z) \right)^{2} d\mu(x)$$

$$= \int_{\mathcal{X}} \sum_{i=1}^{\infty} \left(\frac{\lambda_{i}}{\lambda_{i} + \lambda} \right)^{2} e_{i}^{2}(x) e_{i}^{2}(z) d\mu(x)$$

$$= \sum_{i=1}^{\infty} \left(\frac{\lambda_{i}}{\lambda_{i} + \lambda} \right)^{2} e_{i}^{2}(z).$$
(49)

Since Assumption 3 holds, we have

$$||f||_{L^{\infty}} \le \sum_{i=1}^{\infty} \left(\frac{\lambda_i}{\lambda_i + \lambda}\right)^2 = \mathcal{N}_2(\lambda); \quad ||f||_{L^1} = \sum_{i=1}^{\infty} \left(\frac{\lambda_i}{\lambda_i + \lambda}\right)^2 = \mathcal{N}_2(\lambda).$$

Applying Proposition 34 for \sqrt{f} and noticing that $\|\sqrt{f}\|_{L^{\infty}} = \sqrt{\|f\|_{L^{\infty}}} = \mathcal{N}_2(\lambda)^{\frac{1}{2}}$, we have

$$\frac{1}{2} \left\| \sqrt{f} \right\|_{L^2}^2 - \frac{5\mathcal{N}_2(\lambda)}{3n} \ln \frac{2}{\delta} \le \left\| \sqrt{f} \right\|_{L^2, n}^2 \le \frac{3}{2} \left\| \sqrt{f} \right\|_{L^2}^2 + \frac{5\mathcal{N}_2(\lambda)}{3n} \ln \frac{2}{\delta}, \tag{50}$$

with probability at least $1 - \delta$.

On the one hand, we have

$$\left\| \sqrt{f} \right\|_{L^{2},n}^{2} = \int_{\mathcal{X}} f(\boldsymbol{y}) dP_{n}(\boldsymbol{y}) = \int_{\mathcal{X}} \left[\int_{\mathcal{X}} \left(T_{\lambda}^{-1} k_{\boldsymbol{x}}(\boldsymbol{y}) \right)^{2} d\mu(\boldsymbol{x}) \right] dP_{n}(\boldsymbol{y})$$

$$= \int_{\mathcal{X}} \left[\int_{\mathcal{X}} \left(T_{\lambda}^{-1} k_{\boldsymbol{x}}(\boldsymbol{y}) \right)^{2} dP_{n}(\boldsymbol{y}) \right] d\mu(\boldsymbol{x})$$

$$= \int_{\mathcal{X}} \left\| T_{\lambda}^{-1} k_{\boldsymbol{x}} \right\|_{L^{2},n}^{2} d\mu(\boldsymbol{x}).$$

On the other hand, we have

$$\begin{aligned} \left\| \sqrt{f} \right\|_{L^2}^2 &= \int_{\mathcal{X}} f(\boldsymbol{z}) d\mu(\boldsymbol{z}) \\ &= \int_{\mathcal{X}} \left[\int_{\mathcal{X}} \left(T_{\lambda}^{-1} k_{\boldsymbol{x}}(\boldsymbol{z}) \right)^2 d\mu(\boldsymbol{x}) \right] d\mu(\boldsymbol{z}) \\ &= \int_{\mathcal{X}} \left\| T_{\lambda}^{-1} k_{\boldsymbol{x}} \right\|_{L^2}^2 d\mu(\boldsymbol{x}). \end{aligned}$$

Therefore, (50) implies the desired results.

Approximation A. The proof of Approximation A requires the following proposition, which is a simple but important observation by Li et al. (2023b).

Proposition 11 For any $f, g \in \mathcal{H}$, we have

$$\langle f, g \rangle_{L^2, n} = \langle T_{\boldsymbol{X}} f, g \rangle_{\mathcal{H}} = \langle T_{\boldsymbol{X}}^{1/2} f, T_{\boldsymbol{X}}^{1/2} g \rangle_{\mathcal{H}}.$$
 (51)

Proof Notice that $T_{\mathbf{X}}f = \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{x}_i) k(\mathbf{x}_i, \cdot)$, and thus

$$\langle T_{\boldsymbol{X}} f, g \rangle_{\mathcal{H}} = \frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{x}_i) \langle k(\boldsymbol{x}_i, \cdot), g \rangle_{\mathcal{H}} = \frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{x}_i) g(\boldsymbol{x}_i) = \langle f, g \rangle_{L^2, n}.$$

The second inequality comes from the definition of $T_{\boldsymbol{X}}^{1/2}$.

The following lemma characterizes the magnitude of Approximation A in high probability.

Lemma 12 (Approximation A) Suppose that Assumption 1, 2 and 3 hold. Define $R_1 = R_1(\lambda, X)$ as a function of λ and X:

$$R_{1} := \left| \int_{\mathcal{X}} \left\| T_{\boldsymbol{X}\lambda}^{-1} k_{\boldsymbol{x}} \right\|_{L^{2}, n}^{2} d\mu(\boldsymbol{x}) - \int_{\mathcal{X}} \left\| T_{\lambda}^{-1} k_{\boldsymbol{x}} \right\|_{L^{2}, n}^{2} d\mu(\boldsymbol{x}) \right|.$$
 (52)

Suppose that $\lambda = \lambda(d, n)$ satisfies $\frac{\mathcal{N}_1(\lambda)}{n} \ln n = o(1)$. Then for any fixed $\delta \in (0, 1)$, when n is sufficiently large, with probability at least $1 - \delta$, we have

$$R_1 < 36n^{-1}\mathcal{N}_1(\lambda)^2 \ln n + 12n^{-\frac{1}{2}}\mathcal{N}_1(\lambda)\mathcal{N}_2(\lambda)^{\frac{1}{2}} (\ln n)^{\frac{1}{2}}.$$
 (53)

Proof First, we rewrite R_1 as

$$R_{1} = \left| \int_{\mathcal{X}} \left\| T_{\boldsymbol{X}\lambda}^{-1} k_{\boldsymbol{x}} \right\|_{L^{2}, n}^{2} d\mu(\boldsymbol{x}) - \int_{\mathcal{X}} \left\| T_{\lambda}^{-1} k_{\boldsymbol{x}} \right\|_{L^{2}, n}^{2} d\mu(\boldsymbol{x}) \right|$$

$$\leq \int_{\mathcal{X}} \left\| \left\| T_{\boldsymbol{X}}^{\frac{1}{2}} T_{\boldsymbol{X}\lambda}^{-1} k_{\boldsymbol{x}} \right\|_{\mathcal{H}}^{2} - \left\| T_{\boldsymbol{X}}^{\frac{1}{2}} T_{\lambda}^{-1} k_{\boldsymbol{x}} \right\|_{\mathcal{H}}^{2} d\mu(\boldsymbol{x})$$

$$= \int_{\mathcal{X}} \left\| \left\| T_{\boldsymbol{X}}^{\frac{1}{2}} T_{\boldsymbol{X}\lambda}^{-1} k_{\boldsymbol{x}} \right\|_{\mathcal{H}} - \left\| T_{\boldsymbol{X}}^{\frac{1}{2}} T_{\lambda}^{-1} k_{\boldsymbol{x}} \right\|_{\mathcal{H}} \cdot \left\| T_{\boldsymbol{X}}^{\frac{1}{2}} T_{\boldsymbol{X}\lambda}^{-1} k_{\boldsymbol{x}} \right\|_{\mathcal{H}} + \left\| T_{\boldsymbol{X}}^{\frac{1}{2}} T_{\lambda}^{-1} k_{\boldsymbol{x}} \right\|_{\mathcal{H}} d\mu(\boldsymbol{x}).$$

$$:= \int_{\mathcal{X}} \left| X_{1} - X_{2} \right| \cdot \left| X_{1} + X_{2} \right| d\mu(\boldsymbol{x}).$$

$$(54)$$

where for the second line, we use Proposition 11 to transfer $\|\|_{L^2,n}$ norms into $\|\|_{\mathcal{H}}$.

(I): Given the notations of X_1, X_2 in (54) (both are functions of $\boldsymbol{x}, \boldsymbol{X}$ and λ), we begin to handle $|X_1 - X_2|$:

$$|X_{1} - X_{2}| \leq \left\| T_{\boldsymbol{X}}^{1/2} T_{\boldsymbol{X}\lambda}^{-1} \left(T - T_{\boldsymbol{X}} \right) T_{\lambda}^{-1} k_{\boldsymbol{x}} \right\|_{\mathcal{H}}$$

$$\leq \left\| T_{\boldsymbol{X}}^{1/2} T_{\boldsymbol{X}\lambda}^{-1/2} \right\| \cdot \left\| T_{\boldsymbol{X}\lambda}^{-1/2} T_{\lambda}^{1/2} \right\| \cdot \left\| T_{\lambda}^{-1/2} \left(T - T_{\boldsymbol{X}} \right) T_{\lambda}^{-1/2} \right\| \cdot \left\| T_{\lambda}^{-1/2} k_{\boldsymbol{x}} \right\|_{\mathcal{H}}$$
(55)

(i) Now we bound the third term in (55). Denote $A_i = T_{\lambda}^{-\frac{1}{2}}(T - T_{x_i})T_{\lambda}^{-\frac{1}{2}}$, using Lemma 38, we have

$$||A_i|| = ||T_{\lambda}^{-\frac{1}{2}}TT_{\lambda}^{-\frac{1}{2}}|| + ||T_{\lambda}^{-\frac{1}{2}}T_{\boldsymbol{x}_i}T_{\lambda}^{-\frac{1}{2}}|| \le 2\mathcal{N}_1(\lambda), \quad \mu\text{-a.e. } \boldsymbol{x} \in \mathcal{X}.$$

We use $A \leq B$ to denote that A - B is a positive semi-definite operator. Using the fact that $\mathbb{E}(B - \mathbb{E}B)^2 \leq \mathbb{E}B^2$ for a self-adjoint operator B, we have

$$\mathbb{E}A_i^2 \preceq \mathbb{E}\left[T_{\lambda}^{-\frac{1}{2}}T_{\boldsymbol{x}_i}T_{\lambda}^{-\frac{1}{2}}\right]^2.$$

In addition, Lemma 38 shows that $0 \leq T_{\lambda}^{-\frac{1}{2}} T_{\boldsymbol{x}_i} T_{\lambda}^{-\frac{1}{2}} \leq \mathcal{N}_1(\lambda), \mu\text{-a.e. } \boldsymbol{x} \in \mathcal{X}$. So we have

$$\mathbb{E}A_i^2 \preceq \mathbb{E}\left[T_{\lambda}^{-\frac{1}{2}}T_{\boldsymbol{x}_i}T_{\lambda}^{-\frac{1}{2}}\right]^2 \preceq \mathbb{E}\left[\mathcal{N}_1(\lambda) \cdot T_{\lambda}^{-\frac{1}{2}}T_{\boldsymbol{x}_i}T_{\lambda}^{-\frac{1}{2}}\right] = \mathcal{N}_1(\lambda)T_{\lambda}^{-1}T,$$

Define an operator $V := \mathcal{N}_1(\lambda)T_{\lambda}^{-1}T$, we have

$$||V|| = \mathcal{N}_1(\lambda) \frac{\lambda_1}{\lambda_1 + \lambda} = \mathcal{N}_1(\lambda) \frac{||T||}{||T|| + \lambda} \le \mathcal{N}_1(\lambda);$$

$$\operatorname{tr} V = \mathcal{N}_1(\lambda)^2;$$

$$\frac{\operatorname{tr} V}{||V||} = \frac{\mathcal{N}_1(\lambda)(||T|| + \lambda)}{||T||}.$$

Using Lemma 35 to A_i , V, for any fixed $\delta \in (0,1)$, with probability at least $1-\delta$, we have

$$||T_{\lambda}^{-\frac{1}{2}}(T - T_{\mathbf{X}})T_{\lambda}^{-\frac{1}{2}}|| \le \frac{4\mathcal{N}_{1}(\lambda)}{3n}\beta + \sqrt{\frac{2\mathcal{N}_{1}(\lambda)}{n}\beta},$$

where

$$\beta = \ln \frac{4\mathcal{N}_1(\lambda)(\|T\| + \lambda)}{\delta \|T\|}.$$

Further recall that the condition $\mathcal{N}_1(\lambda) \ln n/n = o(1)$ implies that $\mathcal{N}_1(\lambda) = O(n)$ and thus $\beta = O(\ln n)$, so when n is sufficiently large, we can conclude that

$$||T_{\lambda}^{-\frac{1}{2}}(T - T_{\mathbf{X}})T_{\lambda}^{-\frac{1}{2}}|| \le \sqrt{\frac{2\mathcal{N}_{1}(\lambda)}{n}\beta} \le n^{-\frac{1}{2}}\mathcal{N}_{1}(\lambda)^{\frac{1}{2}}(\ln n)^{\frac{1}{2}}.$$
 (56)

(ii) Next we bound the forth term in (55). Using Lemma 37, we have

$$||T_{\lambda}^{-\frac{1}{2}}k(\boldsymbol{x},\cdot)||_{\mathcal{H}} \leq \mathcal{N}_{1}(\lambda)^{\frac{1}{2}}, \quad \mu\text{-a.e. } \boldsymbol{x} \in \mathcal{X}.$$
 (57)

- (iii) Finally, we bound the first two terms in (55). Since we have assumed $\mathcal{N}_1(\lambda) \ln n/n = o(1)$,
- (56) implies that when n is sufficiently large, we have

$$a := \|T_{\lambda}^{-\frac{1}{2}} (T - T_{\boldsymbol{X}}) T_{\lambda}^{-\frac{1}{2}} \| \le \frac{2}{3}.$$

Therefore, we have

$$\left\| T_{\lambda}^{-\frac{1}{2}} T_{\mathbf{X}\lambda}^{\frac{1}{2}} \right\|^{2} = \left\| T_{\lambda}^{-\frac{1}{2}} T_{\mathbf{X}\lambda} T_{\lambda}^{-\frac{1}{2}} \right\| = \left\| T_{\lambda}^{-\frac{1}{2}} \left(T_{\mathbf{X}} + \lambda \right) T_{\lambda}^{-\frac{1}{2}} \right\|$$

$$= \left\| T_{\lambda}^{-\frac{1}{2}} \left(T_{\mathbf{X}} - T + T + \lambda \right) T_{\lambda}^{-\frac{1}{2}} \right\|$$

$$= \left\| T_{\lambda}^{-\frac{1}{2}} \left(T_{\mathbf{X}} - T \right) T_{\lambda}^{-\frac{1}{2}} + I \right\|$$

$$\leq a + 1 \leq 2; \tag{58}$$

and

$$\left\| T_{\lambda}^{\frac{1}{2}} T_{\mathbf{X}\lambda}^{-\frac{1}{2}} \right\|^{2} = \left\| T_{\lambda}^{\frac{1}{2}} T_{\mathbf{X}\lambda}^{-1} T_{\lambda}^{\frac{1}{2}} \right\| = \left\| \left(T_{\lambda}^{-\frac{1}{2}} T_{\mathbf{X}\lambda} T_{\lambda}^{-\frac{1}{2}} \right)^{-1} \right\|$$

$$= \left\| \left(I - T_{\lambda}^{-\frac{1}{2}} (T_{\mathbf{X}} - T) T_{\lambda}^{-\frac{1}{2}} \right)^{-1} \right\|$$

$$\leq \sum_{k=0}^{\infty} \left\| T_{\lambda}^{-\frac{1}{2}} (T_{\mathbf{X}} - T) T_{\lambda}^{-\frac{1}{2}} \right\|^{k}$$

$$\leq \sum_{k=0}^{\infty} \left(\frac{2}{3} \right)^{k} \leq 3.$$
(59)

Plugging (56), (57), (58) and (59), into (55), when n is sufficiently large, with probability at least $1 - \delta$, we have

$$|X_1 - X_2| \le 6n^{-\frac{1}{2}} \mathcal{N}_1(\lambda) (\ln n)^{\frac{1}{2}}, \quad \mu\text{-a.e. } \mathbf{x} \in \mathcal{X}.$$
 (60)

(II): Further, when n is sufficiently large, with probability at least $1 - \delta$, we also have

$$\int_{\mathcal{X}} X_2 d\mu(\boldsymbol{x}) = \int_{\mathcal{X}} \|T_{\lambda}^{-1} k_{\boldsymbol{x}}\|_{L^2, n} d\mu(\boldsymbol{x})$$

$$\leq \left[\int_{\mathcal{X}} \|T_{\lambda}^{-1} k_{\boldsymbol{x}}\|_{L^2, n}^2 d\mu(\boldsymbol{x}) \right]^{\frac{1}{2}}$$

$$\leq \left(\frac{3}{2} \mathcal{N}_2(\lambda) + R_2 \right)^{\frac{1}{2}}$$

$$\leq (2\mathcal{N}_2(\lambda))^{\frac{1}{2}}, \tag{61}$$

where the third line follows from Lemma 10.

Now we are ready to derive the upper bound of R_1 . Combining the bound (60) and (61), when n is sufficiently large, with probability at least $1 - 2\delta$, we have

$$R_{1} \leq \int_{\mathcal{X}} |X_{1} - X_{2}| \cdot |X_{1} + X_{2}| \, d\mu(\boldsymbol{x})$$

$$= \int_{\mathcal{X}} |X_{1} - X_{2}| \cdot |X_{1} - X_{2} + 2X_{2}| \, d\mu(\boldsymbol{x})$$

$$\leq \int_{\mathcal{X}} |X_{1} - X_{2}|^{2} \, d\mu(\boldsymbol{x}) + \int_{\mathcal{X}} |X_{1} - X_{2}| \cdot 2X_{2} \, d\mu(\boldsymbol{x})$$

$$\leq 36n^{-1} \mathcal{N}_{1}(\lambda)^{2} \ln n + 24n^{-\frac{1}{2}} \mathcal{N}_{1}(\lambda) \mathcal{N}_{2}(\lambda)^{\frac{1}{2}} (\ln n)^{\frac{1}{2}}.$$
(62)

Without loss of generality, we can assume (62) holds with probability at least $1 - \delta$ and we finish the proof.

Final proof of the variance term. Now we are ready to state the theorem about the variance term.

Theorem 13 (Variance term) Suppose that Assumption 1, 2 and 3 hold. If the following approximation conditions hold for some $\lambda = \lambda(d, n) \to 0$:

$$\frac{\mathcal{N}_1(\lambda)}{n}\ln n = o(1); \quad n^{-1}\mathcal{N}_1(\lambda)^2 \ln n = o(\mathcal{N}_2(\lambda)), \tag{63}$$

then we have

$$\mathbf{Var}(\lambda) = \Theta_{\mathbb{P}}\left(\frac{\sigma^2 \mathcal{N}_2(\lambda)}{n}\right). \tag{64}$$

Proof Lemma 9 has shown that

$$\mathbf{Var}(\lambda) = \frac{\sigma^2}{n} \int_{\mathcal{X}} \left\| (T_{\mathbf{X}} + \lambda)^{-1} k_{\mathbf{x}}(\cdot) \right\|_{L^2, n}^2 d\mu(\mathbf{x}).$$

Denote R_1 as in Lemma 12, then conditions (63) and Lemma 12 imply that

$$R_1 = o_{\mathbb{P}} \left(\mathcal{N}_2(\lambda) \right).$$

Further recall that in Lemma 10, we have defined

$$R_2 = \frac{5\mathcal{N}_2(\lambda)}{3n} \ln \frac{2}{\delta}.$$

Then on the one hand, Lemma 10 shows that, for any $\delta \in (0,1)$, when n is sufficiently large, with probability at least $1 - \delta$, we have

$$n\mathbf{Var}(\lambda)/\sigma^{2} = \int_{\mathcal{X}} \|T_{\mathbf{X}\lambda}^{-1} k_{\mathbf{x}}\|_{L^{2},n}^{2} d\mu(\mathbf{x}) \leq \int_{\mathcal{X}} \|T_{\lambda}^{-1} k_{\mathbf{x}}\|_{L^{2},n}^{2} d\mu(\mathbf{x}) + R_{1}$$

$$\leq \frac{3}{2} \int_{\mathcal{X}} \|T_{\lambda}^{-1} k_{\mathbf{x}}\|_{L^{2}}^{2} d\mu(\mathbf{x}) + R_{1} + R_{2}$$

$$= \frac{3}{2} \mathcal{N}_{2}(\lambda) + R_{1} + R_{2},$$

which further implies

$$n\mathbf{Var}(\lambda)/\sigma^2 = O_{\mathbb{P}}\left(\mathcal{N}_2(\lambda)\right). \tag{65}$$

On the other hand, we also have

$$n\mathbf{Var}(\lambda)/\sigma^{2} = \int_{\mathcal{X}} \left\| T_{\mathbf{X}\lambda}^{-1} k_{\mathbf{x}} \right\|_{L^{2},n}^{2} d\mu(\mathbf{x}) \ge \int_{\mathcal{X}} \left\| T_{\lambda}^{-1} k_{\mathbf{x}} \right\|_{L^{2},n}^{2} d\mu(\mathbf{x}) - R_{1}$$

$$\ge \frac{1}{2} \int_{\mathcal{X}} \left\| T_{\lambda}^{-1} k_{\mathbf{x}} \right\|_{L^{2}}^{2} d\mu(\mathbf{x}) - R_{1} - R_{2}$$

$$= \frac{1}{2} \mathcal{N}_{2}(\lambda) - R_{1} - R_{2},$$

which further implies

$$n\mathbf{Var}(\lambda)/\sigma^2 = \Omega_{\mathbb{P}}\left(\mathcal{N}_2(\lambda)\right). \tag{66}$$

Combining (65) and (66), we finish the proof.

A.3 Bias term

In this subsection, our goal is to derive Theorem 16, which shows the upper and lower bounds of bias under some approximation conditions.

The triangle inequality implies that

$$\mathbf{Bias}(\lambda) = \left\| \tilde{f}_{\lambda} - f_{\rho}^* \right\|_{L^2} \ge \left\| f_{\lambda} - f_{\rho}^* \right\|_{L^2} - \left\| \tilde{f}_{\lambda} - f_{\lambda} \right\|_{L^2}, \tag{67}$$

where \tilde{f}_{λ} , f_{λ} are defined as (34) and (36).

The following lemma characterizes the dominant term of $\mathbf{Bias}(\lambda)$.

Lemma 14 Suppose that Assumption 1 and 2 hold. Then for any $\lambda = \lambda(d,n) \to 0$, we have

$$||f_{\lambda} - f_{\rho}^{*}||_{L^{2}} = \mathcal{M}_{2}(\lambda)^{\frac{1}{2}}.$$
 (68)

Proof Recall that we have defined $f_{\rho}^* = \sum_{i=1}^{\infty} f_i e_i(\boldsymbol{x}) \in L^2(\mathcal{X}, \mu)$ and $f_{\lambda} = (T + \lambda)^{-1} S_k^* f_{\rho}^*$. Therefore, we have

$$\begin{aligned} \left\| f_{\lambda} - f_{\rho}^{*} \right\|_{L^{2}}^{2} &= \left\| \sum_{i=1}^{\infty} f_{i} e_{i}(\boldsymbol{x}) - \sum_{i=1}^{\infty} \frac{\lambda_{i}}{\lambda_{i} + \lambda} f_{i} e_{i}(\boldsymbol{x}) \right\|_{L^{2}}^{2} \\ &= \left\| \sum_{i=1}^{\infty} \frac{\lambda}{\lambda_{i} + \lambda} f_{i} e_{i}(\boldsymbol{x}) \right\|_{L^{2}}^{2} \\ &= \sum_{i=1}^{\infty} \left(\frac{\lambda}{\lambda_{i} + \lambda} f_{i} \right)^{2} \\ &= \mathcal{M}_{2}(\lambda). \end{aligned}$$

Our next goal is to prove that second term in (67) is higher order infinitesimal, i.e., $\|\tilde{f}_{\lambda} - f_{\lambda}\|_{L^{2}} = o_{\mathbb{P}}(\mathcal{M}_{2}(\lambda)^{\frac{1}{2}}).$

Lemma 15 Suppose that Assumption 1, 2 and 3 hold. If the following approximation conditions hold for some $\lambda = \lambda(d, n) \to 0$:

$$\frac{\mathcal{N}_1(\lambda)}{n} \ln n = o(1); \quad n^{-1} \mathcal{N}_1(\lambda)^{\frac{1}{2}} \mathcal{M}_1(\lambda) = o\left(\mathcal{M}_2(\lambda)^{\frac{1}{2}}\right), \tag{69}$$

then we have

$$\left\| \tilde{f}_{\lambda} - f_{\lambda} \right\|_{L^{2}} = o_{\mathbb{P}}(\mathcal{M}_{2}(\lambda)^{\frac{1}{2}}). \tag{70}$$

Proof To begin with, be definition, we rewrite (68) as follows

$$\begin{aligned} \left\| \tilde{f}_{\lambda} - f_{\lambda} \right\|_{L_{2}} &= \left\| S_{k} \left(\tilde{f}_{\lambda} - f_{\lambda} \right) \right\|_{L^{2}} \\ &= \left\| S_{k} T_{\lambda}^{-\frac{1}{2}} \cdot T_{\lambda}^{\frac{1}{2}} T_{\boldsymbol{X}\lambda}^{-1} T_{\lambda}^{\frac{1}{2}} \cdot T_{\lambda}^{-\frac{1}{2}} T_{\boldsymbol{X}\lambda} \left(\tilde{f}_{\lambda} - f_{\lambda} \right) \right\|_{L^{2}} \\ &\leq \left\| S_{k} T_{\lambda}^{-\frac{1}{2}} \right\|_{\mathscr{B}(\mathcal{H}, L^{2})} \cdot \left\| T_{\lambda}^{\frac{1}{2}} T_{\boldsymbol{X}\lambda}^{-1} T_{\lambda}^{\frac{1}{2}} \right\|_{\mathscr{B}(\mathcal{H}, \mathcal{H})} \cdot \left\| T_{\lambda}^{-\frac{1}{2}} \left(\tilde{g}_{\boldsymbol{Z}} - T_{\boldsymbol{X}\lambda} f_{\lambda} \right) \right\|_{\mathcal{H}}. \end{aligned}$$
(71)

(i) For any $f \in \mathcal{H}$ and $||f||_{\mathcal{H}} = 1$, suppose that $f = \sum_{i=1}^{\infty} a_i \lambda_i^{1/2} e_i$ satisfying that $\sum_{i=1}^{\infty} a_i^2 = 1$. So for the first term in (71), we have

$$\left\| S_{k} T_{\nu}^{-\frac{1}{2}} \right\|_{\mathscr{B}(\mathcal{H}, L^{2})} = \sup_{\|f\|_{\mathcal{H}} = 1} \left\| S_{k} T_{\nu}^{-\frac{1}{2}} f \right\|_{L^{2}} \\
\leq \sup_{\|f\|_{\mathcal{H}} = 1} \left\| \sum_{i=1}^{\infty} \frac{\lambda_{i}^{\frac{1}{2}}}{(\lambda_{i} + \lambda)^{\frac{1}{2}}} a_{i} e_{i} \right\|_{L^{2}} \\
\leq \sup_{i \geq 1} \frac{\lambda_{i}^{\frac{1}{2}}}{(\lambda_{i} + \lambda)^{\frac{1}{2}}} \cdot \sup_{\|f\|_{\mathcal{H}} = 1} \left\| \sum_{i=1}^{\infty} a_{i} e_{i} \right\|_{L^{2}} \\
\leq 1. \tag{72}$$

(ii) For the second term, since we have assumed $\mathcal{N}_1(\lambda) \ln n/n = o(1)$, for any fixed $\delta \in (0,1)$, when n is sufficiently large, we have proved in (58) and (59) that, with probability at least $1-\delta$

$$\left\| T_{\lambda}^{\frac{1}{2}} T_{\boldsymbol{X}\lambda}^{-1} T_{\lambda}^{\frac{1}{2}} \right\| \le \left\| T_{\lambda}^{\frac{1}{2}} T_{\boldsymbol{X}\lambda}^{-\frac{1}{2}} \right\| \cdot \left\| T_{\boldsymbol{X}\lambda}^{-\frac{1}{2}} T_{\lambda}^{\frac{1}{2}} \right\| \le 3.$$
 (73)

(iii) For the third term in (71), it can be rewritten as

$$\left\| T_{\lambda}^{-\frac{1}{2}} \left(\tilde{g}_{\mathbf{Z}} - T_{\mathbf{X}\lambda} f_{\lambda} \right) \right\|_{\mathcal{H}} = \left\| T_{\lambda}^{-\frac{1}{2}} \left[\left(\tilde{g}_{\mathbf{Z}} - \left(T_{\mathbf{X}} + \lambda + T - T \right) f_{\lambda} \right) \right] \right\|_{\mathcal{H}}$$

$$= \left\| T_{\lambda}^{-\frac{1}{2}} \left[\left(\tilde{g}_{\mathbf{Z}} - T_{\mathbf{X}} f_{\lambda} \right) - \left(T + \lambda \right) f_{\lambda} + T f_{\lambda} \right] \right\|_{\mathcal{H}}$$

$$= \left\| T_{\lambda}^{-\frac{1}{2}} \left[\left(\tilde{g}_{\mathbf{Z}} - T_{\mathbf{X}} f_{\lambda} \right) - \left(g - T f_{\lambda} \right) \right] \right\|_{\mathcal{H}}.$$
(74)

Denote $\xi_i = \xi(\boldsymbol{x}_i) = T_{\lambda}^{-\frac{1}{2}}(K_{\boldsymbol{x}_i}f_{\rho}^*(\boldsymbol{x}_i) - T_{\boldsymbol{x}_i}f_{\lambda})$. To use Bernstein inequality, we need to bound the m-th moment of $\xi(\boldsymbol{x})$:

$$\mathbb{E} \|\xi(\boldsymbol{x})\|_{\mathcal{H}}^{m} = \mathbb{E} \left\| T_{\lambda}^{-\frac{1}{2}} K_{\boldsymbol{x}} (f_{\rho}^{*} - f_{\lambda}(\boldsymbol{x})) \right\|_{\mathcal{H}}^{m}$$

$$\leq \mathbb{E} \left(\left\| T_{\lambda}^{-\frac{1}{2}} k(\boldsymbol{x}, \cdot) \right\|_{\mathcal{H}}^{m} \mathbb{E} \left(\left| (f_{\rho}^{*} - f_{\lambda}(\boldsymbol{x})) \right|^{m} \mid \boldsymbol{x} \right) \right). \tag{75}$$

Note that Lemma 37 shows that

$$\left\| T_{\lambda}^{-\frac{1}{2}}k(\boldsymbol{x},\cdot) \right\|_{\mathcal{H}} \leq \mathcal{N}_{1}(\lambda)^{\frac{1}{2}}, \quad \mu\text{-a.e. } \boldsymbol{x} \in \mathcal{X};$$

By definition of $\mathcal{M}_1(\lambda)$, we also have

$$\left\| f_{\lambda} - f_{\rho}^{*} \right\|_{L^{\infty}} \leq \operatorname{ess sup}_{\boldsymbol{x} \in \mathcal{X}} \left| \sum_{i=1}^{\infty} \frac{\lambda}{\lambda_{i} + \lambda} f_{i} e_{i}(\boldsymbol{x}) \right| = \mathcal{M}_{1}(\lambda).$$
 (76)

In addition, we have proved in Lemma 14 that

$$\mathbb{E}|(f_{\lambda}(\boldsymbol{x}) - f_{\rho}^{*}(\boldsymbol{x}))|^{2} = \mathcal{M}_{2}(\lambda).$$

So we get the upper bound of (75), i.e.,

$$(75) \leq \mathcal{N}_{1}(\lambda)^{\frac{m}{2}} \cdot \|f_{\lambda} - f_{\rho}^{*}\|_{L^{\infty}}^{m-2} \cdot \mathbb{E}|(f_{\lambda}(\boldsymbol{x}) - f_{\rho}^{*}(\boldsymbol{x}))|^{2}$$

$$\leq \mathcal{N}_{1}(\lambda)^{\frac{m}{2}} \mathcal{M}_{1}(\lambda)^{m-2} \mathcal{M}_{2}(\lambda)$$

$$\leq \left(\mathcal{N}_{1}(\lambda)^{\frac{1}{2}} \mathcal{M}_{1}(\lambda)\right)^{m-2} \left(\mathcal{N}_{1}(\lambda)^{\frac{1}{2}} \mathcal{M}_{2}(\lambda)^{\frac{1}{2}}\right)^{2}.$$

Using Lemma 36 with therein notations: $L = \mathcal{N}_1(\lambda)^{\frac{1}{2}}\mathcal{M}_1(\lambda)$ and $\sigma = \mathcal{N}_1(\lambda)^{\frac{1}{2}}\mathcal{M}_2(\lambda)^{\frac{1}{2}}$, for any fixed $\delta \in (0,1)$, with probability at least $1-\delta$, we have

$$(74) \le 4\sqrt{2}\log\frac{2}{\delta}\left(\frac{\mathcal{N}_1(\lambda)^{\frac{1}{2}}\mathcal{M}_1(\lambda)}{n} + \frac{\mathcal{N}_1(\lambda)^{\frac{1}{2}}\mathcal{M}_2(\lambda)^{\frac{1}{2}}}{\sqrt{n}}\right). \tag{77}$$

Since we have assumed $n^{-1}\mathcal{N}_1(\lambda)^{\frac{1}{2}}\mathcal{M}_1(\lambda) = o\left(\mathcal{M}_2(\lambda)^{\frac{1}{2}}\right)$ and $\mathcal{N}_1(\lambda) \ln n/n = o(1)$, (77) further implies

$$\left\| T_{\lambda}^{-\frac{1}{2}} \left(\tilde{g}_{\mathbf{Z}} - T_{\mathbf{X}\lambda} f_{\lambda} \right) \right\|_{\mathcal{H}} = o_{\mathbb{P}} \left(\mathcal{M}_{2}(\lambda)^{\frac{1}{2}} \right). \tag{78}$$

Plugging (72), (73) and (78) into (71), we finish the proof.

Final proof of the bias term. Now we are ready to state the theorem about the bias term.

Theorem 16 Suppose that Assumption 1, 2 and 3 hold. If the following approximation conditions hold for some $\lambda = \lambda(d, n) \to 0$:

$$\frac{\mathcal{N}_1(\lambda)}{n} \ln n = o(1); \quad and \quad n^{-1} \mathcal{N}_1(\lambda)^{\frac{1}{2}} \mathcal{M}_1(\lambda) = o\left(\mathcal{M}_2(\lambda)^{\frac{1}{2}}\right), \tag{79}$$

then we have

$$\mathbf{Bias}^{2}(\lambda) = \Theta_{\mathbb{P}}(\mathcal{M}_{2}(\lambda)). \tag{80}$$

Proof The triangle inequality implies that

$$\mathbf{Bias}(\lambda) = \left\| \tilde{f}_{\lambda} - f_{\rho}^* \right\|_{L^2} \ge \left\| f_{\lambda} - f_{\rho}^* \right\|_{L^2} - \left\| \tilde{f}_{\lambda} - f_{\lambda} \right\|_{L^2},$$

When $\lambda = \lambda(d, n)$ satisfies (79), Lemma 14 and Lemma 15 prove that

$$\left\|f_{\lambda}-f_{
ho}^{*}
ight\|_{L^{2}}=\mathcal{M}_{2}(\lambda)^{rac{1}{2}};\;\;\left\| ilde{f}_{\lambda}-f_{\lambda}
ight\|_{L^{2}}=o_{\mathbb{P}}(\mathcal{M}_{2}(\lambda)^{rac{1}{2}}),$$

which directly prove (80).

A.4 Final proof of Theorem 1

Now we are ready to prove Theorem 1. Note that we have assumed in Theorem 1 that $\lambda = \lambda(d, n) \to 0$ satisfies all the conditions required in Theorem 13 and Theorem 16. Therefore, Theorem 13 and Theorem 16 show that

$$\mathbf{Var}(\lambda) = \Theta_{\mathbb{P}}\left(\frac{\sigma^2 \mathcal{N}_2(\lambda)}{n}\right); \ \ \mathbf{Bias}^2(\lambda) = \Theta_{\mathbb{P}}\left(\mathcal{M}_2(\lambda)\right).$$

Recalling the bias-variance decomposition (38), we finish the proof.

B. Proof of inner product kernel

In this section, we aim to apply Theorem 1 to prove the results in Section 3.2. We will see that the application is nontrivial and is an important contribution of this paper.

We first introduce more necessary preliminaries in Appendix B.1, which is an preparation for subsequent calculations. Next, in order to apply Theorem 1 to get specific convergence rates, we calculate the exact convergence rates of the key quantities therein in Appendix B.2. Finally, we state the proof of Theorem 2 and Theorem 3 in turn in Appendix B.3 and B.4. We will see that there are essential differences in the proof of these two theorems.

B.1 More preliminaries about inner product kernel on the sphere

Suppose that $\mathcal{X} = \mathbb{S}^d$ and μ is the uniform distribution on \mathbb{S}^d . Recall that in Section 3.2, we consider the inner product kernel, i.e., there exists a function $\Phi(t) : [-1,1] \to \mathbb{R}$ such that $k(\boldsymbol{x}, \boldsymbol{x}') = \Phi(\langle \boldsymbol{x}, \boldsymbol{x}' \rangle), \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{S}^d$. Then Mercer's decomposition for the inner product kernel is given in the basis of spherical harmonics:

$$k(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k=0}^{\infty} \mu_k \sum_{l=1}^{N(d,k)} Y_{k,l}(\boldsymbol{x}) Y_{k,l}(\boldsymbol{y}),$$
(81)

where $\{Y_{k,l}\}_{l=1}^{N(d,k)}$ are spherical harmonic polynomials of degree k; μ_k are the eigenvalues with multiplicity N(d,0)=1; $N(d,k)=\frac{2k+d-1}{k}\cdot\frac{(k+d-2)!}{(d-1)!(k-1)!}$, $k=1,2,\cdots$

By known results on spherical harmonics, the eigenvalues μ_k 's have the following explicit expression (Bietti and Mairal, 2019):

$$\mu_k = \frac{\omega_{d-1}}{\omega_d} \int_{-1}^1 \Phi(t) P_k(t) \left(1 - t^2\right)^{(d-2)/2} dt, \tag{82}$$

where P_k is the k-th Legendre polynomial in dimension d+1, ω_d denotes the surface of the sphere \mathbb{S}^d .

Although the above expression of μ_k , N(d,k) are complicated, Lemma 17 \sim 19 (mainly cited from Lu et al. 2023) give concise characterizations of μ_k and N(d,k), which is sufficient for the analysis in this paper.

Lemma 17 Suppose that $k = \{k_d\}_{d=1}^{\infty}$ is a sequence of inner product kernels on the sphere satisfying Assumption 1 and 4. For any fixed integer $p \geq 0$, there exist constants $\mathfrak{C}, \mathfrak{C}_1$ and \mathfrak{C}_2 only depending on p and $\{a_j\}_{j\leq p+1}$, such that for any $d \geq \mathfrak{C}$, we have

$$\mathfrak{C}_1 d^{-k} \le \mu_k \le \mathfrak{C}_2 d^{-k}, \quad k = 0, 1, \dots, p + 1.$$
 (83)

Proof From equation (22) in Ghorbani et al. (2021), for any integer $p \ge 0$, there exist constants \mathfrak{C} only depending on p and $\{a_j\}_{j < p+1}$, such that for any $d \ge \mathfrak{C}$, we have

$$\frac{\Phi^{(k)}(0)}{d^k} \le \mu_k \le \frac{2\Phi^{(k)}(0)}{d^k}, \quad k = 0, 1, \dots, p+1.$$
(84)

Note that for any $k \ge 0$, we have $a_k = \Phi^{(k)}(0)$. Therefore, letting $\mathfrak{C}_1 := \min_{k \le p+1} \{a_k\} > 0$ and $\mathfrak{C}_2 := 2 \max_{k < p+1} \{a_k\} < \infty$, then we finish the proof.

The following property of the eigenvalues $\{\mu_k\}_{k\geq 0}$ indicates that when we consider μ_p with any fixed $p\geq 0$, the subsequent eigenvalues μ_k 's $(k\geq p+1)$ are much smaller than μ_p . The following lemma is the same as Lemma 3.3 in Lu et al. (2023).

Lemma 18 Suppose that $k = \{k_d\}_{d=1}^{\infty}$ is a sequence of inner product kernels on the sphere satisfying Assumption 1 and 4. For any fixed integer $p \geq 0$, there exist constants \mathfrak{C} only depending on p and $\{a_i\}_{i\leq p+1}$, such that for any $d\geq \mathfrak{C}$, we have

$$\mu_k \le \frac{\mathfrak{C}_2}{\mathfrak{C}_1} d^{-1} \mu_p, \quad k = p + 1, p + 2, \cdots$$

where \mathfrak{C}_1 and \mathfrak{C}_2 are constants given in Lemma 17.

Lemma 19 For an integer $k \geq 0$, denote N(d, k) as the multiplicity of the eigenspace corresponding to μ_k in the Mercer's decomposition. For any fixed integer $p \geq 0$, there exist constants \mathfrak{C}_3 , \mathfrak{C}_4 and \mathfrak{C} only depending on p, such that for any $d \geq \mathfrak{C}$, we have

$$\mathfrak{C}_3 d^k \le N(d, k) \le \mathfrak{C}_4 d^k, \quad k = 0, 1, \dots, p + 1.$$
 (85)

Proof When k = 0, we have N(d, 0) = 1, which satisfies (85). When $k \ge 1$, Section 1.6 in Gallier et al. (2020) shows that

$$N(d,k) = \frac{2k+d-1}{k} \cdot \frac{(k+d-2)!}{(d-1)!(k-1)!}.$$

Note that p is fixed and we consider those $k \leq p+1$, (85) follows from detailed calculations using Stirling's approximation. We refer to Lemma B.1 and Lemma D.4 in Lu et al. (2023) for more details.

The following lemma verifies that if we consider the inner product kernel on the sphere, then Assumption 3 naturally holds.

Lemma 20 Suppose that $\mathcal{X} = \mathbb{S}^d$ and μ is the uniform distribution on \mathbb{S}^d . Suppose that k is an inner product kernel, then Assumption 3 holds.

Proof Recall that we have the mercer decomposition (81). Define the sum

$$Z_{k,d}(oldsymbol{x},oldsymbol{y}) = \sum_{l=1}^{N(d,k)} Y_{k,l}(oldsymbol{x}) Y_{k,l}(oldsymbol{y}).$$

Then Dai and Xu (2013, Corollary 1.2.7) shows that $Z_{k,d}(\boldsymbol{x},\boldsymbol{y})$ depends only on $\langle x,y\rangle$ and satisfies

$$|Z_{k,d}(\boldsymbol{x},\boldsymbol{y})| \le Z_{k,d}(\boldsymbol{x},\boldsymbol{x}) = N(d,k), \quad \forall x, y \in \mathbb{S}^d.$$

Therefore, we have

$$\sup_{\boldsymbol{x} \in \mathcal{X}} \sum_{i=1}^{\infty} \left(\frac{\lambda_i}{\lambda_i + \lambda} \right)^2 e_i^2(\boldsymbol{x}) = \sup_{\boldsymbol{x} \in \mathcal{X}} \sum_{k=1}^{\infty} \sum_{l=1}^{N(d,k)} \left(\frac{\mu_k}{\mu_k + \lambda} \right)^2 Y_{k,l}^2(\boldsymbol{x}) = \sum_{k=1}^{\infty} \left(\frac{\mu_k}{\mu_k + \lambda} \right)^2 \sup_{\boldsymbol{x} \in \mathcal{X}} Z_{k,d}(\boldsymbol{x}, \boldsymbol{x})$$
$$= \sum_{k=1}^{\infty} \left(\frac{\mu_k}{\mu_k + \lambda} \right)^2 N(d,k) = \sum_{k=1}^{\infty} \sum_{l=1}^{N(d,k)} \left(\frac{\mu_k}{\mu_k + \lambda} \right)^2$$
$$= \mathcal{N}_2(\lambda).$$

The other equation in Assumption 3 can be proved similarly.

B.2 Calculations of some key quantities

Based on the information of the eigenvalues in the last subsection, this subsection determines the exact convergence rates of the quantities appeared in Theorem 1. These rates will finally determine the convergence rates in Theorem 2 and Theorem 3. Note that we assume d diverges to infinite with n in Theorem 2 and Theorem 3.

Lemma 21 Consider $\mathcal{X} = \mathbb{S}^d$ and the marginal distribution μ to be the uniform distribution. Let $k = \{k_d\}_{d=1}^{\infty}$ be a sequence of inner product kernels on the sphere satisfying Assumption 1 and 4. By choosing $\lambda = d^{-l}$ for some l > 0, we have:

$$\mathcal{N}_1(\lambda) = \Theta\left(\lambda^{-1}\right); \tag{86}$$

If $p \le l \le p+1$ for some $p \in \{0, 1, 2 \cdots\}$, we have

$$\mathcal{N}_2(\lambda) = \Theta\left(d^p + \lambda^{-2}d^{-(p+1)}\right). \tag{87}$$

The notation Θ involves constants only depending on κ and p.

Proof If $p \leq l \leq p+1$ for some $p \in \{0, 1, 2 \cdots\}$, Lemma 17 and Lemma 19 show that there exist constants $\mathfrak{C}, \mathfrak{C}_1, \mathfrak{C}_2, \mathfrak{C}_3$ and \mathfrak{C}_4 only depending on p (recall that we ignore the dependence on $\{a_i\}_{i=0}^{\infty}$), such that for any $d \geq \mathfrak{C}$, we have

$$\mathfrak{C}_{2}^{-1}\mu_{n+1} \le \lambda \le \mathfrak{C}_{1}^{-1}\mu_{n};\tag{88}$$

and for $k = 0, 1, \dots, p + 1$,

$$\mathfrak{C}_1 d^{-k} \le \mu_k \le \mathfrak{C}_2 d^{-k}; \quad \mathfrak{C}_3 d^k \le N(d, k) \le \mathfrak{C}_4 d^k. \tag{89}$$

We first prove (86). On the one hand, for any $d \geq \mathfrak{C}$, we have

$$\mathcal{N}_{1}(\lambda) = \sum_{k=0}^{p} \frac{\mu_{k}}{\mu_{k} + \lambda} N(d, k) + \sum_{k=p+1}^{\infty} \frac{\mu_{k}}{\mu_{k} + \lambda} N(d, k)$$

$$\leq \sum_{k=0}^{p} \frac{\mu_{k}}{\mu_{k}} N(d, k) + \sum_{k=p+1}^{\infty} \frac{\mu_{k}}{\lambda} N(d, k)$$

$$\leq p \mathfrak{C}_{4} d^{p} + \lambda^{-1} \sum_{k=p+1}^{\infty} \mu_{k} N(d, k)$$

$$\leq p \mathfrak{C}_{4} d^{p} + \lambda^{-1} \kappa^{2}$$

$$\lesssim \lambda^{-1}, \tag{90}$$

where we use the fact that $\sum_{k=p+1}^{\infty} \mu_k N(d,k) \leq \sum_{i=1}^{\infty} \lambda_i \leq \sup_{\boldsymbol{x} \in \mathcal{X}} k(x,x) \leq \kappa^2$ for the third inequality.

On the other hand, for any $d \geq \mathfrak{C}$, we have

$$\mathcal{N}_{1}(\lambda) = \sum_{k=0}^{p} \frac{\mu_{k}}{\mu_{k} + \lambda} N(d, k) + \sum_{k=p+1}^{\infty} \frac{\mu_{k}}{\mu_{k} + \lambda} N(d, k)$$

$$\geq \frac{\mu_{p}}{\mu_{p} + \lambda} N(d, p) + \frac{\mu_{p+1}}{\mu_{p+1} + \lambda} N(d, p+1)$$

$$\geq \frac{\mu_{p}}{\mu_{p} + \mathfrak{C}_{1}^{-1} \mu_{p}} N(d, p) + \frac{\mu_{p+1}}{\mathfrak{C}_{2} \lambda + \lambda} N(d, p+1)$$

$$\geq \frac{\mathfrak{C}_{3}}{1 + \mathfrak{C}_{1}^{-1}} d^{p} + \frac{\mathfrak{C}_{1} \mathfrak{C}_{3}}{\mathfrak{C}_{2} + 1} \lambda^{-1}$$

$$\geq \lambda^{-1}. \tag{91}$$

Combining (90) and (91), we finish the proof of (86).

Now we begin to prove (87). On the one hand, for any $d \geq \mathfrak{C}$ we have

$$\mathcal{N}_{2}(\lambda) = \sum_{k=0}^{p} \left(\frac{\mu_{k}}{\mu_{k} + \lambda}\right)^{2} N(d, k) + \sum_{k=p+1}^{\infty} \left(\frac{\mu_{k}}{\mu_{k} + \lambda}\right)^{2} N(d, k)$$

$$\leq \sum_{k=0}^{p} \left(\frac{\mu_{k}}{\mu_{k}}\right)^{2} N(d, k) + \sum_{k=p+1}^{\infty} \left(\frac{\mu_{k}}{\lambda}\right)^{2} N(d, k)$$

$$\leq p \mathfrak{C}_{4} d^{p} + \lambda^{-2} \sum_{k=p+1}^{\infty} \mu_{k}^{2} N(d, k)$$

$$\leq p \mathfrak{C}_{4} d^{p} + \lambda^{-2} \mu_{p+1} \sum_{k=p+1}^{\infty} \mu_{k} N(d, k)$$

$$\leq p \mathfrak{C}_{4} d^{p} + \lambda^{-2} \mathfrak{C}_{2} d^{-(p+1)} \kappa^{2}.$$
(92)

Note that for the forth equation, we use the fact that $\mu_k \leq \mu_{p+1}, \forall k \geq p+1$ (when d is sufficiently large), which can be proved by Lemma 18. On the other hand, for any $d \geq \mathfrak{C}$, we have

$$\mathcal{N}_{2}(\lambda) = \sum_{k=0}^{p} \left(\frac{\mu_{k}}{\mu_{k} + \lambda}\right)^{2} N(d, k) + \sum_{k=p+1}^{\infty} \left(\frac{\mu_{k}}{\mu_{k} + \lambda}\right)^{2} N(d, k)$$

$$\geq \left(\frac{\mu_{k}}{\mu_{p} + \mathfrak{C}_{1}^{-1} \mu_{p}}\right)^{2} N(d, p) + \left(\frac{\mu_{p+1}}{\mathfrak{C}_{2} \lambda + \lambda}\right)^{2} N(d, p+1)$$

$$\geq \frac{\mathfrak{C}_{3}}{\left(1 + \mathfrak{C}_{1}^{-1}\right)^{2}} d^{p} + \lambda^{-2} \left(\frac{\mathfrak{C}_{1}}{\mathfrak{C}_{2} + 1}\right)^{2} \mathfrak{C}_{3} d^{-(p+1)}.$$
(93)

Combining (92) and (93), we prove that $\mathcal{N}_2(\lambda) = \Theta\left(d^p + \lambda^{-2}d^{-(p+1)}\right)$.

Before stating lemmas about $\mathcal{M}_1(\lambda)$, $\mathcal{M}_2(\lambda)$, we first introduce the following useful lemma.

Lemma 22 Consider $\mathcal{X} = \mathbb{S}^d$ and the marginal distribution μ to be the uniform distribution. Let $k = \{k_d\}_{d=1}^{\infty}$ be a sequence of inner product kernels on the sphere satisfying Assumption 1 and 4. Further suppose that Assumption 5 holds for some s > 0. Suppose that α, β are two real numbers such that

$$s + \alpha - \beta < 0; \quad s + \alpha > 0. \tag{94}$$

Then by choosing $\lambda = d^{-l}$ for some $0 < l < \gamma$, if $p \le l \le p+1$ for some $p \in \{0, 1, 2 \cdots\}$, we have

$$\mathcal{M}_{\alpha,\beta}(\lambda) := \sum_{i=1}^{\infty} \frac{\lambda_i^{\alpha}}{(\lambda_i + \lambda)^{\beta}} f_i^2 = \Theta\left(d^{-(s+\alpha-\beta)p} + \lambda^{-\beta} d^{-(p+1)(s+\alpha)}\right). \tag{95}$$

The notation Θ involves constants only depending on s, p, c_0 and R_{γ} , where c_0 and R_{γ} are the constants from Assumption 5.

Proof Similar as the proof of Lemma 21, if $p \le l \le p+1$ for some $p \in \{0, 1, 2 \cdots\}$, there exist constants $\mathfrak{C}, \mathfrak{C}_1, \mathfrak{C}_2, \mathfrak{C}_3$ and \mathfrak{C}_4 only depending on p, such that for any $d \ge \mathfrak{C}$, we have

$$\mathfrak{C}_2^{-1}\mu_{p+1} \le \lambda \le \mathfrak{C}_1^{-1}\mu_p; \tag{96}$$

and for $k = 0, 1, \dots, p + 1$,

$$\mathfrak{C}_1 d^{-k} \le \mu_k \le \mathfrak{C}_2 d^{-k}; \quad \mathfrak{C}_3 d^k \le N(d, k) \le \mathfrak{C}_4 d^k. \tag{97}$$

On the one hand, since (94) holds, for any $d \geq \mathfrak{C}$, we have

$$\mathcal{M}_{\alpha,\beta}(\lambda) = \sum_{k=0}^{\infty} \left(\frac{\mu_k^{s+\alpha}}{(\mu_k + \lambda)^{\beta}} \sum_{i \in \mathcal{I}_{d,k}} \mu_k^{-s} f_i^2 \right)$$

$$\leq \sum_{k=0}^{p} \frac{\mu_k^{s+\alpha}}{(\mu_k + \lambda)^{\beta}} R_{\gamma}^2 + \sum_{k=p+1}^{\infty} \frac{\mu_k^{s+\alpha}}{(\mu_k + \lambda)^{\beta}} \sum_{i \in \mathcal{I}_{d,k}} \mu_k^{-s} f_i^2$$

$$\leq \sum_{k=0}^{p} \mu_k^{s+\alpha-\beta} R_{\gamma}^2 + \sum_{k=p+1}^{\infty} \frac{\mu_k^{s+\alpha}}{\lambda^{\beta}} \sum_{i \in \mathcal{I}_{d,k}} \mu_k^{-s} f_i^2$$

$$\leq p \mu_p^{s+\alpha-\beta} R_{\gamma}^2 + \lambda^{-\beta} \mu_{p+1}^{s+\alpha} \sum_{k=p+1}^{\infty} \sum_{i \in \mathcal{I}_{d,k}} \mu_k^{-s} f_i^2$$

$$\leq p R_{\gamma}^2 \mathfrak{C}_2^{(s+\alpha-\beta)} d^{-(s+\alpha-\beta)p} + \lambda^{-\beta} \mathfrak{C}_2^{s+\alpha} d^{-(p+1)(s+\alpha)} R_{\gamma}^2$$

$$\lesssim d^{-(s+\alpha-\beta)p} + \lambda^{-\beta} d^{-(p+1)(s+\alpha)}. \tag{98}$$

Note that Assumption 5 (a) implies $\sum_{k=0}^{\infty} \sum_{i \in \mathcal{I}_{d,k}} \mu_k^{-s} f_i^2 = \sum_{i=1}^{\infty} \lambda_i^{-s} f_i^2 \leq R_{\gamma}^2$; We also use the fact that $\mu_k \leq \mu_{p+1}, \forall k \geq p+1$, which can be proved by Lemma 18. On the other hand, for any $d \geq \mathfrak{C}$, we have

$$\mathcal{M}_{\alpha,\beta}(\lambda) = \sum_{k=0}^{\infty} \left(\frac{\mu_{k}^{s+\alpha}}{(\mu_{k} + \lambda)^{\beta}} \sum_{i \in \mathcal{I}_{d,k}} \mu_{k}^{-s} f_{i}^{2} \right) \\
\geq \frac{\mu_{p}^{s+\alpha}}{(\mu_{p} + \lambda)^{\beta}} \sum_{i \in \mathcal{I}_{d,p}} \mu_{p}^{-s} f_{i}^{2} + \frac{\mu_{p+1}^{s+\alpha}}{(\mu_{p+1} + \lambda)^{\beta}} \sum_{i \in \mathcal{I}_{d,p+1}} \mu_{p+1}^{-s} f_{i}^{2} \\
\geq \frac{\mu_{p}^{s+\alpha}}{(\mu_{p} + \mathfrak{C}_{1}^{-1} \mu_{p})^{\beta}} \sum_{i \in \mathcal{I}_{d,p}} \mu_{p}^{-s} f_{i}^{2} + \frac{\mu_{p+1}^{s+\alpha}}{(\mathfrak{C}_{2}\lambda + \lambda)^{\beta}} \sum_{i \in \mathcal{I}_{d,p+1}} \mu_{p+1}^{-s} f_{i}^{2} \\
\geq \frac{\mathfrak{C}_{3}^{s+\alpha-\beta}}{(1 + \mathfrak{C}_{1}^{-1})^{\beta}} d^{-(s+\alpha-\beta)p} c_{0} + \lambda^{-\beta} \frac{\mathfrak{C}_{1}^{s+\alpha}}{(\mathfrak{C}_{2} + 1)^{\beta}} d^{-(p+1)(s+\alpha)} c_{0} \\
\geq d^{-(s+\alpha-\beta)p} + \lambda^{-\beta} d^{-(p+1)(s+\alpha)}. \tag{99}$$

We use Assumption 5 (b), i.e., $\sum_{i \in \mathcal{I}_{d,p}} \mu_p^{-s} f_i^2 \ge c_0$ and $\sum_{i \in \mathcal{I}_{d,p+1}} \mu_{p+1}^{-s} f_i^2 \ge c_0$, to obtain the lower bound. Combining (98) and (99), we finish the proof.

Lemma 23 Consider $\mathcal{X} = \mathbb{S}^d$ and the marginal distribution μ to be the uniform distribution. Let $k = \{k_d\}_{d=1}^{\infty}$ be a sequence of inner product kernels on the sphere satisfying Assumption 1 and 4. Further suppose that Assumption 5 holds for some s > 0. Define $\tilde{s} = \min\{s, 2\}$. By choosing $\lambda = d^{-l}$ for some $0 < l < \gamma$, we have: if $p \le l \le p+1$ for some $p \in \{0, 1, 2, \cdots\}$, we have

$$\mathcal{M}_2(\lambda) = \Theta\left(\lambda^2 d^{(2-\tilde{s})p} + d^{-(p+1)\tilde{s}}\right). \tag{100}$$

The notation Θ involves constants only depending on s, p, c_0 and R_{γ} , where c_0 and R_{γ} are the constants from Assumption 5.

Proof When $0 < s \le 2$, $\mathcal{M}_2(\lambda)$ can be viewed as $\lambda^2 \mathcal{M}_{\alpha,\beta}(\lambda)$ in Lemma 22 with $\alpha = 0, \beta = 2$. The conditions (94) are satisfied and Lemma 22 shows that

$$\mathcal{M}_2(\lambda) = \Theta\left(\lambda^2 d^{(2-s)p} + d^{-(p+1)s}\right). \tag{101}$$

When s > 2, without loss of generality, we can assume $\lambda_i \leq 1, \forall i$ and $\lambda \leq 1$. Recall we have proved in Lemma 17 that μ_0 remains as a constant as $d \to \infty$. On the one hand, Assumption 5 (b) also implies $f_1^2 \geq c_0$ without loss of generality, which further implies

$$\mathcal{M}_2(\lambda) = \lambda^2 \sum_{i=1}^{\infty} \frac{f_i^2}{(\lambda_i + \lambda)^2} \ge \frac{1}{4} \lambda^2 \sum_{i=1}^{\infty} f_i^2 \ge \frac{1}{4} \lambda^2 c_0.$$
 (102)

On the other hand, since s > 2, we have

$$\mathcal{M}_{2}(\lambda) = \lambda^{2} \sum_{k=0}^{\infty} \left(\frac{\mu_{k}^{s}}{(\mu_{k} + \lambda)^{2}} \sum_{i \in \mathcal{I}_{d,k}} \mu_{k}^{-s} f_{i}^{2} \right)$$

$$\leq \lambda^{2} \left(\sup_{k \geq 0} \frac{\mu_{k}^{s}}{(\mu_{k} + \lambda)^{2}} \cdot \sum_{k=0}^{\infty} \sum_{i \in \mathcal{I}_{d,k}} \mu_{k}^{-s} f_{i}^{2} \right)$$

$$\leq \lambda^{2} \cdot \sup_{k \geq 0} \frac{\mu_{k}^{s}}{(\mu_{k} + \lambda)^{2}} \cdot R_{\gamma}^{2}$$

$$\leq \lambda^{2} R_{\gamma}^{2}. \tag{103}$$

Further note that since s > 2 and $p \le l \le p+1$, (102) and (103) implies

$$\mathcal{M}_2(\lambda) = \Theta\left(\lambda^2\right) = \Theta\left(\lambda^2 d^{(2-\tilde{s})p} + d^{-(p+1)\tilde{s}}\right).$$

We finish the proof.

The following lemma applies for those $s \geq 1$, which gives an upper bound of $\mathcal{M}_1(\lambda)$.

Lemma 24 Consider $\mathcal{X} = \mathbb{S}^d$ and the marginal distribution μ to be the uniform distribution. Let $k = \{k_d\}_{d=1}^{\infty}$ be a sequence of inner product kernels on the sphere satisfying Assumption 1 and 4. Further suppose that Assumption 5 holds for some $s \geq 1$. Define $\tilde{s} = \min\{s, 2\}$. By choosing $\lambda = d^{-l}$ for some $0 < l < \gamma$, we have: if $p \leq l \leq p+1$ for some $p \in \{0, 1, 2 \cdots\}$, we have

 $\mathcal{M}_1(\lambda) = O\left(\lambda^{\frac{1}{2}} d^{\frac{(2-\tilde{s})p}{2}} + d^{-\frac{(\tilde{s}-1)(p+1)}{2}}\right). \tag{104}$

The notation O involves constants only depending on s, κ, p and R_{γ} , where R_{γ} is the constant from Assumption 5.

Proof First, Cauchy-Schwarz inequality shows that

$$\mathcal{M}_{1}(\lambda)^{2} = \operatorname{ess sup} \left| \sum_{i=1}^{\infty} \left(\frac{\lambda}{\lambda_{i} + \lambda} f_{i} e_{i}(\boldsymbol{x}) \right) \right|^{2}$$

$$\leq \left(\sum_{i=1}^{\infty} \frac{\lambda^{2} \lambda_{i}^{-1}}{\lambda_{i} + \lambda} f_{i}^{2} \right) \cdot \operatorname{ess sup}_{\boldsymbol{x} \in \mathcal{X}} \sum_{i=1}^{\infty} \left(\frac{\lambda_{i}}{\lambda_{i} + \lambda} e_{i}(\boldsymbol{x})^{2} \right)$$

$$\leq \left(\sum_{i=1}^{\infty} \frac{\lambda^{2} \lambda_{i}^{-1}}{\lambda_{i} + \lambda} f_{i}^{2} \right) \cdot \sum_{i=1}^{\infty} \left(\frac{\lambda_{i}}{\lambda_{i} + \lambda} \right)$$

$$:= \mathcal{Q}_{1}(\lambda) \cdot \mathcal{N}_{1}(\lambda), \tag{105}$$

where we use (6) in Assumption 3 for the second inequality.

When $1 \le s \le 2$, $Q_1(\lambda)$ defined above can be viewed as $\lambda^2 \mathcal{M}_{\alpha,\beta}(\lambda)$ in Lemma 22 with $\alpha = -1, \beta = 1$. In addition, the conditions (94) are satisfied, thus Lemma 22 shows that

$$Q_1(\lambda) = \Theta\left(\lambda^2 d^{(2-s)p} + \lambda d^{-(s-1)(p+1)}\right). \tag{106}$$

Since we assume the kernel to be bounded in Assumption 1, we can assume $\lambda_i \leq 1, \forall i$ and $\lambda \leq 1$ without loss of generality. When s > 2, on the one hand, Assumption 5 (b) also implies

$$Q_1(\lambda) = \lambda^2 \sum_{i=1}^{\infty} \frac{f_i^2}{(\lambda_i + \lambda) \lambda_i} \ge \frac{1}{2} \lambda^2 \sum_{i=1}^{\infty} f_i^2 \ge \frac{1}{2} \lambda^2 c_0.$$
 (107)

On the other hand, since s > 2, we have

$$\mathcal{Q}_{1}(\lambda) = \lambda^{2} \sum_{k=0}^{\infty} \left(\frac{\mu_{k}^{s-1}}{\mu_{k} + \lambda} \sum_{i \in \mathcal{I}_{d,k}} \mu_{k}^{-s} f_{i}^{2} \right) \\
\leq \lambda^{2} \left(\sup_{k \geq 0} \frac{\mu_{k}^{s-1}}{\mu_{k} + \lambda} \cdot \sum_{k=0}^{\infty} \sum_{i \in \mathcal{I}_{d,k}} \mu_{k}^{-s} f_{i}^{2} \right) \\
\leq \lambda^{2} \cdot \frac{\mu_{k}^{s-1}}{\mu_{k} + \lambda} \cdot R_{\gamma}^{2} \\
\leq \lambda^{2} R_{\gamma}^{2}. \tag{108}$$

Further note that since s > 2 and $p \le l \le p + 1$, (107) and (108) implies

$$Q_1(\lambda) = \Theta(\lambda^2) = \Theta(\lambda^2 d^{(2-\tilde{s})p} + \lambda d^{-(\tilde{s}-1)(p+1)}).$$

Therefore, we have $Q_1(\lambda) = \Theta\left(\lambda^2 d^{(2-\tilde{s})p} + \lambda d^{-(\tilde{s}-1)(p+1)}\right)$ for any s > 0.

Further recalling that Lemma 21 proves $\mathcal{N}_1(\lambda) = \Theta(\lambda^{-1})$, use (105) and we finish the proof.

When 0 < s < 1, the following lemma gives an upper bound of $||f_{\lambda}||_{L^{\infty}}$.

Lemma 25 Consider $\mathcal{X} = \mathbb{S}^d$ and the marginal distribution μ to be the uniform distribution. Let $k = \{k_d\}_{d=1}^{\infty}$ be a sequence of inner product kernels on the sphere satisfying Assumption 1 and 4. Further suppose that Assumption 5 holds for some 0 < s < 1. Recall the definition of f_{λ} in (36). By choosing $\lambda = d^{-l}$ for some $0 < l < \gamma$, we have: if $p \le l \le p+1$ for some $p \in \{0, 1, 2 \cdots\}$, we have

$$||f_{\lambda}||_{L^{\infty}} = O\left(d^{\frac{(1-s)p}{2}} + \lambda^{-1}d^{-\frac{(1+s)(p+1)}{2}}\right).$$
 (109)

The notation O involves constants only depending on s, κ, p , and R_{γ} , where R_{γ} is the constant from Assumption 5.

Proof We need the following fact: For any $f \in \mathcal{H}$, since Assumption 1 holds, we have

$$||f_{\lambda}||_{L^{\infty}} \leq \sup_{\boldsymbol{x} \in \mathcal{X}} |f(\boldsymbol{x})| = \sup_{\boldsymbol{x} \in \mathcal{X}} \langle f(\cdot), k(\boldsymbol{x}, \cdot) \rangle_{\mathcal{H}}$$

$$\leq \sup_{\boldsymbol{x} \in \mathcal{X}} ||k(\boldsymbol{x}, \cdot)||_{\mathcal{H}} \cdot ||f||_{\mathcal{H}}$$

$$< \kappa^{2} ||f||_{\mathcal{H}}.$$

Therefore, by definition and notice that $f_{\lambda} \in \mathcal{H}$, we have

$$||f_{\lambda}||_{L^{\infty}}^{2} \le \kappa^{4} ||f_{\lambda}||_{\mathcal{H}}^{2} = \kappa^{4} \sum_{i=1}^{\infty} \frac{\lambda_{i}}{(\lambda_{i} + \lambda)^{2}} f_{i}^{2} := \mathcal{Q}_{2}(\lambda).$$
 (110)

Since 0 < s < 1, $\mathcal{Q}_2(\lambda)$ can be viewed as $\kappa^4 \mathcal{M}_{\alpha,\beta}(\lambda)$ in Lemma 22 with $\alpha = 1, \beta = 2$ and the conditions (94) are satisfied. Thus Lemma 22 shows that

$$Q_2(\lambda) = \Theta\left(\lambda^2 d^{(1-s)p} + \lambda^{-s} d^{-(1+s)(p+1)}\right).$$

Taking the square root, we finish the proof.

B.3 Proof of Theorem 2

In the last subsection, we have calculated the exact convergence rates of $\mathcal{N}_1(\lambda)$, $\mathcal{N}_2(\lambda)$, $\mathcal{M}_1(\lambda)$, $\mathcal{M}_2(\lambda)$ when $\lambda = d^{-l}$ for some $0 < l < \gamma$. Note that we have proved in Lemma 20 that Assumption 3 naturally holds for inner product kernel on the sphere. Now we are ready to apply Theorem 1 to prove Theorem 2. The proof mainly consists of 3 steps:

- (1) For specific range of $\gamma > 0$, we use Lemma 21 and Lemma 23 to derive the scale of λ_{balance} or l_{balance} such that $\mathcal{N}_2(\lambda)/n$ and $\mathcal{M}_2(\lambda)$ are balanced.
- (2) We check that the conditions (7) required in Theorem 1 are satisfied for $\lambda \gtrsim \lambda_{\text{balance}}$ (or $\lambda = d^{-l}, l \leq l_{\text{balance}}$).
- (3) Using the monotonicity of $\mathbf{Var}(\lambda)$ with respect to λ , we demonstrate that $\lambda = \lambda_{\text{balance}}$ is the best choice of regularization parameter, i.e., the generalization error of KRR estimator is the smallest when $\lambda = \lambda_{\text{balance}}$. That is to say, the convergence rate of the generalization error can not be faster than the rate when choosing $\lambda = \lambda_{\text{balance}}$.

Note that we expect $\mathbf{Bias}^2(\lambda) = \Theta_{\mathbb{P}}(\mathcal{M}_2(\lambda))$ and $\mathbf{Var}(\lambda) = \Theta_{\mathbb{P}}(\mathcal{N}_2(\lambda)/n)$. Step 1 actually indicates the regularization such that the bias and variance are balanced. Together with Theorem 13 and 16, Step 2 further verifies that $\mathbf{Bias}^2(\lambda) = \Theta_{\mathbb{P}}(\mathcal{M}_2(\lambda))$ and $\mathbf{Var}(\lambda) = \Theta_{\mathbb{P}}(\mathcal{N}_2(\lambda)/n)$ indeed hold for those $\lambda \gtrsim \lambda_{\text{balance}}$. Thus they are indeed balanced under the choice of $\lambda = \lambda_{\text{balance}}$.

Final proof of Theorem 2. In the following of the proof, we omit the dependence of constants on $s, \sigma, \gamma, c_0, \kappa, c_1$ and c_2 .

Step 1: Note that we assume $s \ge 1$ in this theorem and $\lambda = d^{-l}$, $0 < l < \gamma$. For specific range of γ , we discuss the range of l_{balance} . Recall that we define $\tilde{s} = \min\{s, 2\}$.

• When $l \in (p, p + \frac{1}{2}]$ for some integer $p \ge 0$, Lemma 21 and Lemma 23 show that

$$\frac{\mathcal{N}_2(\lambda)}{n} \asymp d^{p-\gamma}; \quad \mathcal{M}_2(\lambda) \asymp d^{-2l+(2-\tilde{s})p},$$

thus we have

$$l_{\text{balance}} = \frac{\gamma + p - p\tilde{s}}{2}.\tag{111}$$

Further, letting $l_{\text{balance}} = \frac{\gamma + p - p\tilde{s}}{2} \in (p, p + \frac{1}{2}]$, we have

$$\gamma \in (p + p\tilde{s}, p + p\tilde{s} + 1]. \tag{112}$$

• When $l \in (p + \frac{1}{2}, p + \frac{\tilde{s}}{2}]$, Lemma 21 and Lemma 23 show that

$$\frac{\mathcal{N}_2(\lambda)}{n} \asymp d^{2l-p-1-\gamma}; \quad \mathcal{M}_2(\lambda) \asymp d^{-2l+(2-\tilde{s})p},$$

thus we have

$$l_{\text{balance}} = \frac{\gamma + 3p - p\tilde{s} + 1}{4}.\tag{113}$$

Further, letting $l_{\text{balance}} = \frac{\gamma + 3p - p\tilde{s} + 1}{4} \in (p + \frac{1}{2}, p + \frac{\tilde{s}}{2}]$, we have

$$\gamma \in (p + p\tilde{s} + 1, p + p\tilde{s} + 2\tilde{s} - 1].$$
 (114)

• When $l \in (p + \frac{\tilde{s}}{2}, p + 1]$, Lemma 21 and Lemma 23 show that

$$\frac{\mathcal{N}_2(\lambda)}{n} \asymp d^{2l-p-1-\gamma}; \quad \mathcal{M}_2(\lambda) \asymp d^{-(p+1)\tilde{s}},$$

thus we have

$$l_{\text{balance}} = \frac{\gamma + (p+1)(1-\tilde{s})}{2}.\tag{115}$$

Further, letting $l_{\text{balance}} \in (p + \frac{\tilde{s}}{2}, p + 1]$, we have

$$\gamma \in (p + p\tilde{s} + 2\tilde{s} - 1, (p+1) + (p+1)\tilde{s}].$$
 (116)

Step 2: In order to apply Theorem 13 and Theorem 16 so that we know the exact convergence rates of $Var(\lambda_{balance})$ and $Bias^2(\lambda_{balance})$, we first check the approximation conditions (63) and (69), or equivalently conditions (7), hold for $l = l_{balance}$. Recall that we have calculated the convergence rates of $\mathcal{N}_1(\lambda)$ and $\mathcal{M}_1(\lambda)$ in Lemma 21 and Lemma 24.

- When $\gamma \in (p + p\tilde{s}, p + p\tilde{s} + 1]$: recall that $l_{\text{balance}} = \frac{\gamma + p p\tilde{s}}{2} \in (p, p + \frac{1}{2}]$.
 - (i) The first condition in (7) is equivalent to

$$\frac{\gamma + p - p\tilde{s}}{2} < \gamma \iff \gamma > p - p\tilde{s},$$

which naturally holds for all $\gamma \in (p + p\tilde{s}, p + p\tilde{s} + 1]$.

(ii) The second condition in (7) is equivalent to

$$d^{-\gamma} \cdot d^{\gamma + p - p\tilde{s}} \cdot \gamma \ln d \ll d^p \iff p - p\tilde{s} < p$$

which naturally holds for all $\gamma \in (p + p\tilde{s}, p + p\tilde{s} + 1]$ and $p \neq 0$. When p = 0, we actually need to choose $\lambda_{\text{balance}} = d^{-l_{\text{balance}}} \cdot \ln d$ and the second condition will hold.

(iii) The third condition in (7) is equivalent to

$$d^{-\gamma} \cdot d^{\frac{\gamma + p - p\tilde{s}}{4}} \cdot \left[d^{-\frac{\gamma + p - p\tilde{s}}{4} + \frac{(2 - \tilde{s})p}{2}} + d^{-\frac{(\tilde{s} - 1)(p + 1)}{2}} \right] \ll d^{-\frac{1}{2}(\gamma + p - p\tilde{s}) + \frac{(2 - \tilde{s})p}{2}}$$

 \iff

$$\gamma > p - p\tilde{s}; \quad \gamma > p - 3p\tilde{s} - 2\tilde{s} + 2,$$

which naturally holds for all $\gamma \in (p + p\tilde{s}, p + p\tilde{s} + 1]$ and $p \neq 0$. In addition, one can also check that the third condition in (7) holds when p = 0 and $\lambda_{\text{balance}} = d^{-l_{\text{balance}}} \cdot \ln d$.

- When $\gamma \in (p + p\tilde{s} + 1, p + p\tilde{s} + 2\tilde{s} 1]$: recall that $l_{\text{balance}} = \frac{\gamma + 3p p\tilde{s} + 1}{4} \in (p + \frac{1}{2}, p + \frac{\tilde{s}}{2}]$.
 - (i) The first condition in (7) is equivalent to

$$\frac{\gamma + 3p - p\tilde{s} + 1}{4} < \gamma \iff \gamma > p - \frac{p\tilde{s}}{3} + \frac{1}{3},$$

which naturally holds for all $\gamma \in (p + p\tilde{s} + 1, p + p\tilde{s} + 2\tilde{s} - 1]$.

(ii) The second condition in (7) is equivalent to

$$d^{-\gamma} \cdot d^{\frac{\gamma+3p-p\tilde{s}+1}{2}} \cdot \gamma \ln d \ll d^{\frac{\gamma+3p-p\tilde{s}+1}{2}-p-1} \iff \gamma > p+1,$$

which naturally holds for all $\gamma \in (p + p\tilde{s} + 1, p + p\tilde{s} + 2\tilde{s} - 1]$.

(iii) The third condition in (7) is equivalent to

$$d^{-\gamma} \cdot d^{\frac{\gamma + p - p\tilde{s}}{8}} \cdot \left[d^{-\frac{\gamma + p - p\tilde{s}}{8} + \frac{(2 - \tilde{s})p}{2}} + d^{-\frac{(\tilde{s} - 1)(p + 1)}{2}} \right] \ll d^{-\frac{1}{4}(\gamma + p - p\tilde{s}) + \frac{(2 - \tilde{s})p}{2}}$$

 \iff

$$\gamma > p - \frac{p\tilde{s}}{3} + \frac{1}{3}; \quad \gamma > p - \frac{3p\tilde{s}}{5} - \frac{4\tilde{s}}{5} + \frac{7}{5},$$

which naturally holds for all $\gamma \in (p + p\tilde{s} + 1, p + p\tilde{s} + 2\tilde{s} - 1]$.

- When $\gamma \in (p+p\tilde{s}+2\tilde{s}-1,(p+1)+(p+1)\tilde{s}]$: recall that $l_{\text{balance}} = \frac{\gamma+(p+1)(1-\tilde{s})}{2} \in (p+\frac{\tilde{s}}{2},p+1.$
 - (i) The first condition in (7) is equivalent to

$$\frac{\gamma + (p+1)(1-\tilde{s})}{2} < \gamma \iff \gamma > p - p\tilde{s} + 1 - \tilde{s},$$

which naturally holds for all $\gamma \in (p + p\tilde{s} + 2\tilde{s} - 1, (p+1) + (p+1)\tilde{s}]$.

(ii) The second condition in (7) is equivalent to

$$d^{-\gamma} \cdot d^{\gamma + (p+1)(1-\tilde{s})} \cdot \gamma \ln d \ll d^{\gamma + (p+1)(1-\tilde{s})-p-1} \iff \gamma > p+1,$$

which naturally holds for all $\gamma \in (p + p\tilde{s} + 2\tilde{s} - 1, (p + 1) + (p + 1)\tilde{s}]$.

(iii) The third condition in (7) is equivalent to

$$d^{-\gamma} \cdot d^{\frac{\gamma + (p+1)(1-\tilde{s})}{4}} \cdot \left[d^{-\frac{\gamma + (p+1)(1-\tilde{s})}{2} + \frac{(2-\tilde{s})p}{2}} + d^{-\frac{(\tilde{s}-1)(p+1)}{2}} \right] \ll d^{-\frac{(p+1)\tilde{s}}{2}}$$

$$\gamma>p+\frac{\tilde{s}}{2}; \quad \gamma>p-\frac{p\tilde{s}}{3}+\frac{\tilde{s}}{3}+\frac{1}{3},$$

which naturally holds for all $\gamma \in (p + p\tilde{s} + 2\tilde{s} - 1, (p+1) + (p+1)\tilde{s}]$.

Up to now, we have verified conditions (7) for $l = l_{\text{balance}}$. Furthermore, simple calculation shows that the order of

$$\frac{\mathcal{N}_1(\lambda)}{n}; \quad n^{-1}\mathcal{N}_1(\lambda)^2 \mathcal{N}_2(\lambda)^{-1}; \quad n^{-1}\mathcal{N}_1(\lambda)^{\frac{1}{2}} \mathcal{M}_1(\lambda) \mathcal{M}_2(\lambda)^{-\frac{1}{2}}$$
(117)

are all non-decreasing with respect to l, where we choose $\lambda = d^{-l}$. Therefore, the above results indicate that conditions (7) holds for all $l \leq l_{\text{balance}}$.

Step 3: In step 2, on the one hand, we prove that by choosing $\lambda = \lambda_{\text{balance}}$,

$$\mathbb{E}\left[\left\|\hat{f}_{\lambda_{\text{balance}}} - f_{\rho}^{*}\right\|_{L^{2}}^{2} \mid \boldsymbol{X}\right] = \Theta_{\mathbb{P}}\left(\frac{\sigma^{2}\mathcal{N}_{2}(\lambda_{\text{balance}})}{n} + \mathcal{M}_{2}(\lambda_{\text{balance}})\right). \tag{118}$$

On the other hand, we also prove that by choosing $\lambda \gtrsim \lambda_{\text{balance}}$,

$$\mathbb{E}\left[\left\|\hat{f}_{\lambda_{\text{balance}}} - f_{\rho}^{*}\right\|_{L^{2}}^{2} \mid \boldsymbol{X}\right] = \Omega_{\mathbb{P}}\left(\frac{\sigma^{2}\mathcal{N}_{2}(\lambda_{\text{balance}})}{n} + \mathcal{M}_{2}(\lambda_{\text{balance}})\right). \tag{119}$$

In the following, we handle those $\lambda \lesssim \lambda_{\text{balance}}$. Recall that we have shown in the proof of Lemma 9 that

$$\mathbf{Var}(\lambda) = \frac{\sigma^2}{n^2} \int_{\mathcal{X}} \mathbb{K}(\boldsymbol{x}, \boldsymbol{X}) (\boldsymbol{K} + \lambda)^{-2} \mathbb{K}(\boldsymbol{X}, \boldsymbol{x}) \, d\mu(\boldsymbol{x}).$$

One simple but critical observation from Li et al. (2023b) is that

$$(\mathbf{K} + \lambda_1)^{-2} \succeq (\mathbf{K} + \lambda_2)^{-2}$$
, if $\lambda_1 \leq \lambda_2$,

where \succeq represents the partial order of positive semi-definite matrices, and thus for $\lambda_1 \le \lambda_2 > 0$, we have

$$\mathbf{Var}(\lambda_1) \geq \mathbf{Var}(\lambda_2).$$

Also note that by the definition of l_{balance} , we actually have

$$\mathbb{E}\left[\left\|\hat{f}_{\lambda_{\text{balance}}} - f_{\rho}^{*}\right\|_{L^{2}}^{2} \mid \boldsymbol{X}\right] = \Theta_{\mathbb{P}}\left(\mathbf{Var}(\lambda_{\text{balance}})\right). \tag{120}$$

Therefore, for those $\lambda \lesssim \lambda_{\text{balance}}$, we have

$$\mathbb{E}\left[\left\|\hat{f}_{\lambda} - f_{\rho}^{*}\right\|_{L^{2}}^{2} \mid \boldsymbol{X}\right] \geq \operatorname{Var}(\lambda) \geq \operatorname{Var}(\lambda_{\operatorname{balance}}) \approx \mathbb{E}\left[\left\|\hat{f}_{\lambda_{\operatorname{balance}}} - f_{\rho}^{*}\right\|_{L^{2}}^{2} \mid \boldsymbol{X}\right]. \quad (121)$$

To sum up, (118), (119) and (121) show that by choosing $\lambda = \lambda_{\text{balance}}$ as in step 1, we obtain the convergence rates of KRR estimator under the best regularization. Using Lemma 23 to calculate the rate of $\mathcal{M}_2(\lambda_{\text{balance}})$, we finish the proof.

B.4 Proof of Theorem 3

Recall that in the proof of Theorem 16, we use $\mathcal{M}_1(\lambda)$ to bound $\|f_{\lambda} - f_{\rho}^*\|_{L^{\infty}}$ and show that $\|\tilde{f}_{\lambda} - f_{\lambda}\|_{L^2} = o_{\mathbb{P}}(\mathcal{M}_2(\lambda)^{\frac{1}{2}})$. Unfortunately, the calculation of $\mathcal{M}_1(\lambda)$ in Lemma 24 only holds for $s \geq 1$ and it could be infinite when s < 1. Extensive literature then assume $\|f_{\rho}^*\|_{L^{\infty}}$ to be bounded and use $\|f_{\lambda}\|_{L^{\infty}} + \|f_{\rho}^*\|_{L^{\infty}}$ to bound $\|f_{\lambda} - f_{\rho}^*\|_{L^{\infty}}$. When the dimension d is fixed, Zhang et al. (2023a) first use a truncation method together with the L^q -embedding property of $[\mathcal{H}]^s$ to remove the boundedness assumption when s < 1. We will see in the proof of Theorem 3 that this technique still works in the large-dimensional setting.

To be specific, when we further assume $f_{\rho}^* \in [\mathcal{H}]^s$, we have the following Theorem which is a refined version of Lemma 15.

Lemma 26 Suppose that Assumption 1, 2 and 3 hold. Further suppose that Assumption 5 holds for some 0 < s < 1. If the following approximation conditions hold for some $\lambda = \lambda(d, n) \to 0$:

$$\frac{\mathcal{N}_1(\lambda)}{n} \ln n = o(1); \quad n^{-1} \mathcal{N}_1(\lambda)^{\frac{1}{2}} \|f_{\lambda}\|_{L^{\infty}} = o\left(\mathcal{M}_2(\lambda)^{\frac{1}{2}}\right); \tag{122}$$

and there exists $\epsilon > 0$, such that

$$n^{-1}\mathcal{N}_1(\lambda)^{\frac{1}{2}}n^{\frac{1-s}{2}+\epsilon} = o\left(\mathcal{M}_2(\lambda)^{\frac{1}{2}}\right),\tag{123}$$

then we have

$$\left\| \tilde{f}_{\lambda} - f_{\lambda} \right\|_{L^{2}} = o_{\mathbb{P}}(\mathcal{M}_{2}(\lambda)^{\frac{1}{2}}), \tag{124}$$

where the notation $o_{\mathbb{P}}$ involves constants only depending on s and κ .

Proof Recall the decomposition (71) in the proof of Lemma 15. The first two terms in (71) can be handled without any difference. Our goal here is to prove the following equation and we will finish the proof:

$$\left\| T_{\lambda}^{-\frac{1}{2}} \left(\tilde{g}_{\mathbf{Z}} - T_{\mathbf{X}\lambda} f_{\lambda} \right) \right\|_{\mathcal{H}} = o_{\mathbb{P}} \left(\mathcal{M}_{2}(\lambda)^{\frac{1}{2}} \right). \tag{125}$$

Similar as (74), we rewrite the left hand side of (125) as

$$\left\| T_{\lambda}^{-\frac{1}{2}} \left(\tilde{g}_{\mathbf{Z}} - T_{\mathbf{X}\lambda} f_{\lambda} \right) \right\|_{\mathcal{H}} = \left\| T_{\lambda}^{-\frac{1}{2}} \left[\left(\tilde{g}_{\mathbf{Z}} - T_{\mathbf{X}} f_{\lambda} \right) - \left(g - T f_{\lambda} \right) \right] \right\|_{\mathcal{H}}.$$
 (126)

Denote $\xi_i = \xi(\boldsymbol{x}_i) = T_{\lambda}^{-\frac{1}{2}}(K_{\boldsymbol{x}_i}f_{\rho}^*(\boldsymbol{x}_i) - T_{\boldsymbol{x}_i}f_{\lambda})$. Further consider the subset $\Omega_1 = \{\boldsymbol{x} \in \mathcal{X} : |f_{\rho}^*(\boldsymbol{x})| \leq t\}$ and $\Omega_2 = \mathcal{X} \setminus \Omega_1$, where t will be chosen appropriately later. Decompose ξ_i as $\xi_i I_{\boldsymbol{x}_i \in \Omega_1} + \xi_i I_{\boldsymbol{x}_i \in \Omega_2}$ and we have the following decomposition of (126):

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \xi_{i} - \mathbb{E} \xi_{x} \right\|_{\mathcal{H}} \leq \left\| \frac{1}{n} \sum_{i=1}^{n} \xi_{i} I_{\boldsymbol{x}_{i} \in \Omega_{1}} - \mathbb{E} \xi_{\boldsymbol{x}} I_{\boldsymbol{x} \in \Omega_{1}} \right\|_{\mathcal{H}} + \left\| \frac{1}{n} \sum_{i=1}^{n} \xi_{i} I_{\boldsymbol{x}_{i} \in \Omega_{2}} \right\|_{\mathcal{H}} + \left\| \mathbb{E} \xi_{\boldsymbol{x}} I_{\boldsymbol{x} \in \Omega_{2}} \right\|_{\mathcal{H}}$$

$$:= I + II + III. \tag{127}$$

Next we choose $t=n^{\frac{1-s}{2}+\epsilon_t}, q=\frac{2}{1-s}-\epsilon_q$ such that

$$\epsilon_t < \epsilon; \text{ and } \frac{1-s}{2} + \epsilon_t > 1/\left(\frac{2}{1-s} - \epsilon_q\right),$$
(128)

where ϵ is given in (123). Then we can bound the three terms in (127) as follows:

(i) For the first term in (127), denoted as I, notice that

$$\|(f_{\lambda} - f_{\rho}^{*}) I_{x_{i} \in \Omega_{1}}\|_{L^{\infty}} \le \|f_{\lambda}\|_{L^{\infty}} + n^{\frac{1-s}{2} + \epsilon_{t}}.$$
 (129)

Imitating the procedure (iii) in the proof of Lemma 15 and using (122), (123), we have

$$I = o_{\mathbb{P}} \left(\mathcal{M}_2(\lambda)^{\frac{1}{2}} \right). \tag{130}$$

(ii) For the second term in (127), denoted as II. Since $q = \frac{2}{1-s} - \epsilon_q < \frac{2}{1-s}$, Lemma 42 shows that,

$$[\mathcal{H}]^s \hookrightarrow L^q(\mathcal{X}, \mu),$$
 (131)

with embedding norm less than a constant $C_{s,\kappa}$. Then Assumption 5 (a) implies that there exists $0 < C_q < \infty$ only depending on γ, s and κ such that $\|f_{\rho}^*\|_{L^q(\mathcal{X},\mu)} \leq C_q$. Using the Markov inequality, we have

$$P(\boldsymbol{x} \in \Omega_2) = P(|f_{\rho}^*(\boldsymbol{x})| > t) \le \frac{\mathbb{E}|f_{\rho}^*(\boldsymbol{x})|^q}{t^q} \le \frac{(C_q)^q}{t^q}.$$

Further, since (128) guarantees $t^q \gg n$, we have

$$\tau_{n} := P\left(\text{II} \ge \mathcal{M}_{2}(\lambda)^{\frac{1}{2}}\right) \le P\left(\exists \boldsymbol{x}_{i} \text{ s.t. } \boldsymbol{x}_{i} \in \Omega_{2},\right) = 1 - P\left(\boldsymbol{x}_{i} \notin \Omega_{2}, \forall \boldsymbol{x}_{i}, i = 1, 2, \cdots, n\right)$$

$$= 1 - P\left(\boldsymbol{x} \notin \Omega_{2}\right)^{n}$$

$$= 1 - P\left(|f_{\rho}^{*}(\boldsymbol{x})| \le t\right)^{n}$$

$$\le 1 - \left(1 - \frac{(C_{q})^{q}}{t^{q}}\right)^{n} \to 0.$$
(132)

(iii) For the third term in (127), denoted as III. Since Lemma 37 implies that $||T_{\lambda}^{-\frac{1}{2}}k(\boldsymbol{x},\cdot)||_{\mathcal{H}} \leq \mathcal{N}_{1}(\lambda)^{\frac{1}{2}}, \mu\text{-a.e. } \boldsymbol{x} \in \mathcal{X}$, so

$$\begin{aligned}
&\text{III} \leq \mathbb{E} \| \xi_{\boldsymbol{x}} I_{\boldsymbol{x} \in \Omega_{2}} \|_{\mathcal{H}} \leq \mathbb{E} \Big[\| T_{\lambda}^{-\frac{1}{2}} k(\boldsymbol{x}, \cdot) \|_{\mathcal{H}} \cdot \left| \left(f_{\rho}^{*} - f_{\lambda}(\boldsymbol{x}) \right) I_{\boldsymbol{x} \in \Omega_{2}} \right| \\
&\leq \mathcal{N}_{1}(\lambda)^{\frac{1}{2}} \mathbb{E} \left| \left(f_{\rho}^{*} - f_{\lambda}(\boldsymbol{x}) \right) I_{\boldsymbol{x} \in \Omega_{2}} \right| \\
&\leq \mathcal{N}_{1}(\lambda)^{\frac{1}{2}} \| f_{\rho}^{*} - f_{\lambda} \|_{L^{2}}^{\frac{1}{2}} \cdot P\left(\boldsymbol{x} \in \Omega_{2} \right)^{\frac{1}{2}} \\
&\leq \mathcal{N}_{1}(\lambda)^{\frac{1}{2}} \mathcal{M}_{2}(\lambda)^{\frac{1}{2}} t^{-\frac{q}{2}},
\end{aligned} \tag{133}$$

where we use Cauchy-Schwarz inequality for the third inequality and (14) for the forth inequality. Recalling that the choices of t,q satisfy $t^{-q} \ll n^{-1}$ and we have assumed $\mathcal{N}_1(\lambda) \ln n/n = o(1)$, we have

$$III = o\left(\mathcal{M}_2(\lambda)^{\frac{1}{2}}\right). \tag{134}$$

Plugging (130), (132) and (134) into (127), we finish the proof.

Based on Lemma 26 and Lemma 14, we have the following theorem about the exact rate of bias term when 0 < s < 1.

Theorem 27 Suppose that Assumption 1, 2 and 3 hold. Further suppose that Assumption 5 holds for some 0 < s < 1. If the following approximation conditions hold for some $\lambda = \lambda(d, n) \to 0$:

$$\frac{\mathcal{N}_1(\lambda)}{n} \ln n = o(1); \quad n^{-1} \mathcal{N}_1(\lambda)^{\frac{1}{2}} \|f_{\lambda}\|_{L^{\infty}} = o\left(\mathcal{M}_2(\lambda)^{\frac{1}{2}}\right); \tag{135}$$

and there exists $\epsilon > 0$, such that

$$n^{-1}\mathcal{N}_1(\lambda)^{\frac{1}{2}}n^{\frac{1-s}{2}+\epsilon} = o\left(\mathcal{M}_2(\lambda)^{\frac{1}{2}}\right),\tag{136}$$

then we have

$$\mathbf{Bias}^{2}(\lambda) = \Theta_{\mathbb{P}}\left(\mathcal{M}_{2}(\lambda)\right),\tag{137}$$

where the notation $\Theta_{\mathbb{P}}$ involves constants only depending on s and κ .

Proof The triangle inequality implies that

$$\mathbf{Bias}(\lambda) = \left\| \tilde{f}_{\lambda} - f_{\rho}^* \right\|_{L^2} \ge \left\| f_{\lambda} - f_{\rho}^* \right\|_{L^2} - \left\| \tilde{f}_{\lambda} - f_{\lambda} \right\|_{L^2},$$

When $\lambda = \lambda(d, n)$ satisfies (135) and (136), Lemma 14 and Lemma 26 prove that

$$\left\|f_{\lambda}-f_{\rho}^{*}
ight\|_{L^{2}}=\mathcal{M}_{2}(\lambda)^{rac{1}{2}};\;\;\left\| ilde{f}_{\lambda}-f_{\lambda}
ight\|_{L^{2}}=o_{\mathbb{P}}(\mathcal{M}_{2}(\lambda)^{rac{1}{2}}),$$

which directly prove (137).

Now we are ready to prove Theorem 3. Since we do not claim that the regularization choice in Theorem 3 is the best, we only need the first two steps in the proof of Theorem 2.

Final proof of Theorem 3. In the following of the proof, we omit the dependence of constants on $s, \sigma, \gamma, c_0, \kappa, c_1$ and c_2 .

Step 1: Note that we assume 0 < s < 1 in this theorem and $\lambda = d^{-l}$, $0 < l < \gamma$. For specific range of γ , we discuss the range of l_{balance} .

• When $l \in (p, p + \frac{s}{2}]$ for some integer $p \ge 0$, Lemma 21 and Lemma 23 show that

$$\frac{\mathcal{N}_2(\lambda)}{n} \asymp d^{p-\gamma}; \quad \mathcal{M}_2(\lambda) \asymp d^{-2l+(2-s)p},$$

thus we have

$$l_{\text{balance}} = \frac{\gamma + p - ps}{2}.\tag{138}$$

Further, letting $l_{\text{balance}} = \frac{\gamma + p - ps}{2} \in (p, p + \frac{s}{2}]$, we have

$$\gamma \in (p + ps, p + ps + s]. \tag{139}$$

• When $l \in (p + \frac{s}{2}, p + \frac{1}{2}]$, Lemma 21 and Lemma 23 show that

$$\frac{\mathcal{N}_2(\lambda)}{n} \asymp d^{p-\gamma}; \quad \mathcal{M}_2(\lambda) \asymp d^{-(p+1)s},$$

thus the above two terms are equal if and only if

$$\gamma = p + ps + s. \tag{140}$$

• When $l \in (p + \frac{1}{2}, p + 1]$, Lemma 21 and Lemma 23 show that

$$\frac{\mathcal{N}_2(\lambda)}{n} \asymp d^{2l-p-1-\gamma}; \quad \mathcal{M}_2(\lambda) \asymp d^{-(p+1)s},$$

thus we have

$$l_{\text{balance}} = \frac{\gamma + (p+1)(1-s)}{2}.$$
 (141)

Further, letting $l_{\text{balance}} \in (p + \frac{s}{2}, p + 1]$, we have

$$\gamma \in (p + ps + s, (p+1) + (p+1)s]. \tag{142}$$

Note that the present result is different from the result of the Step 1 in the proof of Theorem (2). There are only two intervals of γ , i.e.,

$$\gamma \in (p + ps, p + ps + s]; \text{ and } \gamma \in (p + ps + s, (p + 1) + (p + 1)s].$$

It is worth mentioning that in the second interval of γ , we can actually choose $\lambda = d^{-l}, \forall l \in [p + \frac{s}{2}, l_{\text{balance}}]$ and we have

$$\frac{\mathcal{N}_2(\lambda)}{n} \lesssim \mathcal{M}_2(\lambda); \quad \mathcal{M}_2(\lambda) = \mathcal{M}_2(\lambda_{\text{balance}}). \tag{143}$$

That is to say, we can choose smaller l and the rate of $\mathcal{N}_2(\lambda)/n + \mathcal{M}_2(\lambda)$ will remain unchanged. We have shown in (117) and the discussion below it that the approximation conditions are easier to satisfied for smaller l. Therefore, in the following of the proof, we define

$$l_{\text{opt}} = p + \frac{s}{2}$$
, when $\gamma \in (p + ps + s, (p+1) + (p+1)s]$, (144)

and verify the approximation conditions for $\lambda_{\rm opt}=d^{-l_{\rm opt}}.$ For consistency of notation, we also define

$$l_{\text{opt}} = \frac{\gamma + p - ps}{2}$$
, when $\gamma \in (p + ps, p + ps + s]$. (145)

Step 2: In order to apply Theorem 13 and Theorem 27 so that we know the exact convergence rates of $Var(\lambda_{opt})$ and $Bias^2(\lambda_{opt})$, we first check the approximation conditions (63), (135) and (136) hold for $l = l_{opt}$. We first list all the approximation conditions below:

$$\frac{\mathcal{N}_{1}(\lambda)}{n} \ln n = o(1); \quad n^{-1} \mathcal{N}_{1}(\lambda)^{2} \ln n = o(\mathcal{N}_{2}(\lambda));
n^{-1} \mathcal{N}_{1}(\lambda)^{\frac{1}{2}} \|f_{\lambda}\|_{L^{\infty}} = o\left(\mathcal{M}_{2}(\lambda)^{\frac{1}{2}}\right); n^{-1} \mathcal{N}_{1}(\lambda)^{\frac{1}{2}} n^{\frac{1-s}{2}+\epsilon} = o\left(\mathcal{M}_{2}(\lambda)^{\frac{1}{2}}\right).$$
(146)

Recall that we have calculated the convergence rates of $\mathcal{N}_1(\lambda)$ and $||f_{\lambda}||_{L^{\infty}}$ in Lemma 21 and Lemma 25.

- When $\gamma \in (p+ps, p+ps+s]$: recall that $l_{\text{opt}} = \frac{\gamma + p ps}{2} \in (p, p + \frac{s}{2}]$.
 - (i) The first condition in (146) is equivalent to

$$\frac{\gamma + p - ps}{2} < \gamma \iff \gamma > p - ps,$$

which naturally holds for all $\gamma \in (p + ps, p + ps + s]$.

(ii) The second condition in (146) is equivalent to

$$d^{-\gamma} \cdot d^{\gamma + p - ps} \cdot \gamma \ln d \ll d^p \iff p - ps < p,$$

which naturally holds for all $\gamma \in (p + ps, p + ps + s]$ and $p \neq 0$. When p = 0, we actually need to choose $\lambda_{\text{opt}} = d^{-l_{\text{opt}}} \cdot \ln d$ and the second condition will hold.

(iii) The third condition in (146) is equivalent to

$$d^{-\gamma} \cdot d^{\frac{\gamma + p - ps}{4}} \cdot \left[d^{\frac{p}{2} - \frac{ps}{2}} + d^{\frac{\gamma + p - ps}{2} - \frac{(1 + s)(p + 1)}{2}} \right] \ll d^{-\frac{1}{2}(\gamma + p - ps) + \frac{(2 - s)p}{2}}$$

 \iff

$$\gamma > p - 3ps; \quad \gamma$$

which naturally holds for all $\gamma \in (p+ps, p+ps+s]$ and $p \neq 0$. In addition, one can also check that the third condition in (146) holds when p=0 and $\lambda_{\rm opt}=d^{-l_{\rm opt}}\cdot \ln d$. (iv) The forth condition in (146) is equivalent to

$$d^{-\gamma} \cdot d^{\frac{\gamma + p - ps}{4}} \cdot d^{\frac{\gamma}{2} - \frac{\gamma s}{2}} \ll d^{-\frac{1}{2}(\gamma + p - ps) + \frac{(2 - s)p}{2}}$$

 \iff

$$(1-2s)\gamma$$

If $\frac{1}{2} < s < 1$, (147) naturally holds for all $\gamma \in (p + ps, p + ps + s]$ and $p \neq 0$. In addition, one can also check that the forth condition in (146) holds when p = 0 and $\lambda_{\text{opt}} = d^{-l_{\text{opt}}} \cdot \ln d$.

If $0 < s \le \frac{1}{2}$, (147) only holds for $\gamma \in (p + ps, p + ps + s]$ and $p \ne 0$. That is to say, we can not verify (147) holds for

$$\gamma \in (0, s], \quad 0 < s < \frac{1}{2}.$$
 (148)

- When $\gamma \in (p+ps+s,(p+1)+(p+1)s]$: recall that $l_{\text{opt}}=p+\frac{s}{2}$.
 - (i) The first condition in (146) is equivalent to

$$p + \frac{s}{2} < \gamma,$$

which naturally holds for all $\gamma \in (p + ps + s, (p + 1) + (p + 1)s]$.

(ii) The second condition in (146) is equivalent to

$$d^{-\gamma} \cdot d^{2p+s} \cdot \gamma \ln d \ll d^p \iff \gamma > p+s,$$

which naturally holds for all $\gamma \in (p + ps + s, (p + 1) + (p + 1)s]$.

(iii) The third condition in (146) is equivalent to

$$d^{-\gamma} \cdot d^{\frac{p}{2} + \frac{s}{4}} \cdot \left[d^{\frac{p}{2} - \frac{ps}{2}} + d^{p + \frac{s}{2} - \frac{(1+s)(p+1)}{2}} \right] \ll d^{-\frac{(p+1)s}{2}}$$

 \iff

$$\gamma>p+\frac{3s}{4}; \quad \gamma>p+\frac{3s}{4}-\frac{1}{2},$$

which naturally holds for all $\gamma \in (p + ps + s, (p + 1) + (p + 1)s]$.

(iv) The forth condition in (146) is equivalent to

$$d^{-\gamma} \cdot d^{\frac{p}{2} + \frac{s}{4}} \cdot d^{\frac{\gamma}{2} - \frac{\gamma s}{2}} \ll d^{-\frac{(p+1)s}{2}}$$

 \iff

$$\gamma > \frac{2p + 3s + 2ps}{2(s+1)}. (149)$$

If 1/2 < s < 1, (149) naturally holds for all $\gamma \in (p+ps+s,(p+1)+(p+1)s]$ and $p \neq 0$. In addition, one can also check that the forth condition in (146) holds when p=0 and $\lambda_{\rm opt} = d^{-l_{\rm opt}} \cdot \ln d$.

If $0 < s \le 1/2$, (149) only holds for $\gamma \in (p + ps + s, (p + 1) + (p + 1)s]$ and $p \ne 0$. That is to say, we can not verify (147) holds for

$$\gamma \in \left(0, \frac{3s}{2(s+1)}\right], \quad 0 < s < \frac{1}{2}.$$
(150)

Up to now, we have verified the approximation conditions (146) for

$$\forall \gamma > 0, \text{ if } \frac{1}{2} < s < 1;$$
 (151)

and

$$\forall \gamma > \frac{3s}{2(s+1)}, \text{ if } 0 < s \le \frac{1}{2}.$$
 (152)

Using Lemma 23 to calculate the rate of $\mathcal{M}_2(\lambda_{\text{opt}})$, we finish the proof.

C. Proof of Minimax lower bound

C.1 More preliminaries about minimax lower bound

Let's first introduce several concepts about minimax lower bound which can be frequently found in related literature Yang and Barron (1999); Lu et al. (2023), etc..

Suppose that (\mathcal{Z}, d) is a topological space with a compatible loss function d, which are mappings from $\mathcal{Z} \times \mathcal{Z}$ to $\mathbb{R}_{\geq 0}$ with d(f, f) = 0 and d(f, f') > 0 for $f \neq f'$. We call such a loss function a *distance*. We introduce the packing entropy and covering entropy below:

Definition 28 (Packing entropy) A finite set $N_{\epsilon} \subset \mathcal{Z}$ is said to be an ϵ -packing set in \mathcal{Z} with separation $\epsilon > 0$, if for any $f, f' \in N_{\epsilon}, f \neq f'$, we have $d(f, f') > \epsilon$. The logarithm of the maximum cardinality of ϵ -packing set is called the ϵ -packing entropy or Kolmogorov capacity of \mathcal{Z} with distance d and is denoted by $M_d(\epsilon, \mathcal{Z})$.

Definition 29 (Covering entropy) A set $G_{\epsilon} \subset \mathcal{Z}$ is said to be an ϵ -net for \mathcal{Z} if for any $\tilde{f} \in \mathcal{Z}$, there exists an $f_0 \in G_{\epsilon}$ such that $d(\tilde{f}, f_0) \leq \epsilon$. The logarithm of the minimum cardinality of ϵ -net is called the ϵ -covering entropy of \mathcal{Z} and is denoted by $V_d(\epsilon, \mathcal{Z})$.

Let $\mathcal{B} = \{f \in \mathcal{H}, \|f\|_{[\mathcal{H}]^s} \leq R_{\gamma}\}$, where R_{γ} is the constant from Assumption 5. Without loss of generality, we can consider \mathcal{B} be the unit ball in $[\mathcal{H}]^s$. Let $M_2(\epsilon, \mathcal{B})$ be the ϵ -packing entropy of $(\mathcal{B}, d^2 = \|\cdot\|_{L^2}^2)$ and $V_2(\epsilon, \mathcal{B})$ be the ϵ -covering entropy of $(\mathcal{B}, d^2 = \|\cdot\|_{L^2}^2)$. Recalling that μ is the marginal distribution on \mathcal{X} , we further define

$$\mathcal{D} = \left\{ \rho_f \mid \text{ joint distribution of } (y, \boldsymbol{x}) \text{ where } \boldsymbol{x} \sim \mu, y = f(\boldsymbol{x}) + \epsilon, \epsilon \sim N(0, \sigma^2), f \in \mathcal{B} \right\},$$

and let $V_K(\epsilon, \mathcal{D})$ be the ϵ -covering entropy of $(\mathcal{D}, d^2 = \text{KL divergence})$. It is easy to see that \mathcal{D} is an subset of \mathcal{P} which is defined in Theorem 5, i.e., $\mathcal{D} \subset \mathcal{P}$.

The following lemmas give useful characterizations of $M_2(\epsilon, \mathcal{B}), V_2(\epsilon, \mathcal{B})$ and $V_K(\epsilon, \mathcal{D})$. We refer to Lemma A.5, Lemma A.7 and Lemma A.8 in Lu et al. (2023) for their proofs.

Lemma 30 For any $\epsilon > 0$, we have $M_2(2\epsilon, \mathcal{B}) \leq V_2(\epsilon, \mathcal{B}) \leq M_2(\epsilon, \mathcal{B})$.

Lemma 31
$$V_2(\epsilon, \mathcal{B}) = V_K\left(\frac{\epsilon}{\sqrt{2}\sigma}, \mathcal{D}\right)$$
.

Lemma 32 Let $\{\lambda_j\}_{j=1}^{\infty}$ be the eigenvalues of \mathcal{H} . For any $\epsilon > 0$, let $K(\epsilon) = \frac{1}{2} \sum_{j:\lambda_j^s > \epsilon^2} \ln\left(\lambda_j^s/\epsilon^2\right)$.

We have

$$V_2(6\epsilon, \mathcal{B}) \le K(\epsilon) \le V_2(\epsilon, \mathcal{B}).$$
 (153)

The following important lemma is a modification of Theorem 1 and Corollary 1 in Yang and Barron (1999). We refer to Lemma 4.1 in Lu et al. (2023) for the proof.

Lemma 33 Let $\mathfrak{c} \in (0,1)$ be a constant only depending on c_1 , c_2 , and γ , where c_1, c_2 are the constants given in Theorem 5. For any $0 < \tilde{\epsilon}_1, \tilde{\epsilon}_2 < \infty$ only depending on n, d, $\{\lambda_j\}$, c_1 , c_2 , and γ and satisfying

$$\frac{V_K(\tilde{\epsilon}_2, \mathcal{D}) + n\tilde{\epsilon}_2^2 + \ln 2}{V_2(\tilde{\epsilon}_1, \mathcal{B})} \le \mathfrak{c},\tag{154}$$

we have

$$\min_{\hat{f}} \max_{\rho_{f^*} \in \mathcal{D}} \mathbb{E}_{(\boldsymbol{X}, \mathbf{y}) \sim \rho_{f^*}^{\otimes n}} \left\| \hat{f} - f^* \right\|_{L^2}^2 \ge \frac{1 - \mathfrak{c}}{4} \tilde{\epsilon}_1^2.$$
 (155)

C.2 Proof of Theorem 5

Now we are ready to use the lemmas in the last subsection to prove Theorem 5. The proof is divided into two parts, dealing with the two cases of the interval in which γ falls into.

Proof of Theorem 5 (i). In this case, we have assumed $\gamma \in (p+ps, p+ps+s]$ for some integer $p \geq 0$. Let $\tilde{\epsilon}_2 = C_2 d^{-(\gamma-p)}$, where we will choose the constant C_2 later. Note that we have $\gamma - p \in (ps, (p+1)s]$. Lemma 17 implies that we can choose C_2 only depending on p (ignoring the dependence on $\{a_j\}_{j=0}^{\infty}$) such that for any $d \geq \mathfrak{C}$ (\mathfrak{C} is a constant only depending on ϵ , s and p), we have

$$\mu_{p+1}^s < \tilde{\epsilon}_2^2 < \mu_p^s. \tag{156}$$

Next we can choose $\tilde{\epsilon}_1 = d^{-(\gamma - p + \epsilon)}$, where ϵ can be any positive real number. Since $\gamma - p + \epsilon > ps$, when $d \geq \mathfrak{C}$, where \mathfrak{C} is a constant only depending on ϵ, s and p, we have

$$\tilde{\epsilon}_1^2 < \mu_p^s. \tag{157}$$

Therefore, using Lemma 32 and Lemma 17, for any $d \geq \mathfrak{C}$, we have

$$V_{2}(\tilde{\epsilon}_{1}, \mathcal{B}) \geq K(\tilde{\epsilon}_{1}) \geq \frac{1}{2}N(d, p)\ln\left(\frac{\mu_{p}^{s}}{\tilde{\epsilon}_{1}^{2}}\right)$$

$$\geq \frac{1}{2}N(d, p)\ln\left(\frac{\mathfrak{C}_{1}d^{-ps}}{d^{-(\gamma-p+\epsilon)}}\right)$$

$$= \frac{1}{2}N(d, p)\left(\ln\mathfrak{C}_{1} + (\gamma - p + \epsilon - ps)\ln d\right). \tag{158}$$

In addition, using Lemma 17 and Lemma 18, we have the following claim.

Claim 1 Suppose that $\gamma \in (p+ps, p+ps+s]$ for some integer $p \geq 0$. Let $\tilde{\epsilon}_2^2$ be defined as above. For any $\epsilon_0 > 0$, there exists a sufficiently large constant \mathfrak{C} only depending on s, p and ϵ_0 , such that for any $d \geq \mathfrak{C}$, we have

$$K\left(\sqrt{2}\sigma\tilde{\epsilon}_2/6\right) \le (1+\epsilon_0)\frac{1}{2}N(d,p)\log\left(\frac{18\mu_p^s}{\sigma^2\tilde{\epsilon}_2^2}\right).$$

Therefore, for any $d \geq \mathfrak{C}$, where \mathfrak{C} is a constant only depending on s and p, we have

$$V_{K}(\tilde{\epsilon}_{2}, \mathcal{D}) = V_{2}\left(\sqrt{2}\sigma\tilde{\epsilon}_{2}, \mathcal{B}\right) \leq K\left(\frac{\sqrt{2}\sigma\tilde{\epsilon}_{2}}{6}\right)$$

$$\leq (1 + \epsilon_{0})\frac{1}{2}N(d, p)\ln\left(\frac{18\mu_{p}^{s}}{\sigma^{2}\tilde{\epsilon}_{2}^{2}}\right)$$

$$\leq (1 + \epsilon_{0})\frac{1}{2}N(d, p)\ln\left(\frac{18\mathfrak{C}_{2}d^{-ps}}{\sigma^{2}C_{2}d^{-(\gamma - p)}}\right)$$

$$= (1 + \epsilon_{0})\frac{1}{2}N(d, p)\left(\ln\frac{18\mathfrak{C}_{2}}{\sigma^{2}C_{2}} + (\gamma - p - ps)\ln d\right), \tag{159}$$

where we use Lemma 31 and Lemma 32 for the first line and use Lemma 17 for the third line.

Using (158) and (159), also recalling that we assume $c_1 d^{\gamma} \leq n \leq c_2 d^{\gamma}$, we have

$$\frac{V_K(\tilde{\epsilon}_2, \mathcal{D}) + n\tilde{\epsilon}_2^2 + \ln 2}{V_2(\tilde{\epsilon}_1, \mathcal{B})} \le \frac{(1 + \epsilon_0) \frac{1}{2} N(d, p) \left(\ln \frac{18\mathfrak{C}_2}{\sigma^2 C_2} + (\gamma - p - ps) \ln d \right) + c_2 d^{\gamma} \cdot C_2 d^{-(\gamma - p)} + \ln 2}{\frac{1}{2} N(d, p) \left(\ln \mathfrak{C}_1 + (\gamma - p + \epsilon - ps) \ln d \right)}.$$
(160)

Recalling that Lemma 19 shows $\mathfrak{C}_3 d^p \leq N(d,p) \leq \mathfrak{C}_4 d^p$ when $d \geq \mathfrak{C}$, the dominant terms in (160) are:

$$\frac{\frac{1}{2}\left(1+\epsilon_0\right)\left(\gamma-p-ps\right)N(d,p)\ln d}{\frac{1}{2}\left(\gamma-p+\epsilon-ps\right)N(d,p)\ln d}.$$
(161)

Therefore, for any $\epsilon > 0$, we can choose ϵ_0 small enough such that

$$\frac{V_K(\tilde{\epsilon}_2, \mathcal{D}) + n\tilde{\epsilon}_2^2 + \ln 2}{V_2(\tilde{\epsilon}_1, \mathcal{B})} \le (160) := \mathfrak{c} < 1. \tag{162}$$

Then using Lemma 33, we have

$$\min_{\hat{f}} \max_{\rho_{f^*} \in \mathcal{D}} \mathbb{E}_{(\boldsymbol{X}, \mathbf{y}) \sim \rho_{f^*}^{\otimes n}} \left\| \hat{f} - f^* \right\|_{L^2}^2 \geq \frac{1 - \mathfrak{c}}{4} \tilde{\epsilon}_1^2 = \frac{1 - \mathfrak{c}}{4} d^{-(\gamma - p - \epsilon)}.$$

Further recalling that $\mathcal{D} \subset \mathcal{P}$, we have

$$\min_{\hat{f}} \max_{\rho \in \mathcal{P}} \mathbb{E}_{(\boldsymbol{X}, \mathbf{y}) \sim \rho^{\otimes n}} \left\| \hat{f} - f_{\rho}^{*} \right\|_{L^{2}}^{2} \ge \min_{\hat{f}} \max_{\rho_{f^{*}} \in \mathcal{D}} \mathbb{E}_{(\boldsymbol{X}, \mathbf{y}) \sim \rho_{f^{*}}^{\otimes n}} \left\| \hat{f} - f^{*} \right\|_{L^{2}}^{2} \ge \frac{1 - \mathfrak{c}}{4} d^{-(\gamma - p - \epsilon)}. \tag{163}$$

We finis the proof of Theorem 5 (i).

Proof of Theorem 5 (ii). In this case, we have assumed $\gamma \in (p+ps+s,(p+1)+(p+1)s]$ for some integer $p \geq 0$. Let $\tilde{\epsilon}_2 = C_2 d^{-(p+1)s} \ln d$, where we will choose the constant C_2 later. Then Lemma 17 implies that there exists a constant $\mathfrak C$ only depending on s and p such that for any $d \geq \mathfrak C$, we have

$$\mu_{p+1}^s < \tilde{\epsilon}_2^2 < \mu_p^s. \tag{164}$$

Next we can choose $\tilde{\epsilon}_1 = C_1 d^{-(p+1)s}$. Using Lemma 17, we can choose $C_1 < \mathfrak{C}_1^s$, where \mathfrak{C}_1 is the constant in Lemma 17, such that for any $d \geq \mathfrak{C}$, where \mathfrak{C} is a constant only depending on s and p, we have

$$\hat{\epsilon}_1^2 < \mu_{p+1}^s. \tag{165}$$

Therefore, using Lemma 32 and Lemma 17, for any $d \geq \mathfrak{C}$, we have

$$V_{2}(\tilde{\epsilon}_{1}, \mathcal{B}) \geq K(\tilde{\epsilon}_{1}) \geq \frac{1}{2}N(d, p+1)\ln\left(\frac{\mu_{p+1}^{s}}{\tilde{\epsilon}_{1}^{2}}\right)$$

$$\geq \frac{1}{2}N(d, p+1)\ln\left(\frac{\mathfrak{C}_{1}d^{-(p+1)s}}{C_{1}d^{-(p+1)s}}\right)$$

$$= \frac{1}{2}N(d, p+1)\ln\frac{\mathfrak{C}_{1}}{C_{1}}.$$
(166)

In addition, using Lemma 17 and Lemma 18, we have the following claim.

Claim 2 Suppose that $\gamma \in (p+ps+s,(p+1)+(p+1)s]$ for some integer $p \geq 0$. Let $\tilde{\epsilon}_2^2$ be defined as above. For any $\epsilon_0 > 0$, there exists a sufficiently large constant \mathfrak{C} only depending on s, p and ϵ_0 , such that for any $d \geq \mathfrak{C}$, we have

$$K\left(\sqrt{2}\sigma\tilde{\epsilon}_2/6\right) \le (1+\epsilon_0)\frac{1}{2}N(d,p)\log\left(\frac{18\mu_p^s}{\sigma^2\tilde{\epsilon}_2^2}\right).$$

Therefore, for any $d \geq \mathfrak{C}$, where \mathfrak{C} is a constant only depending on s, p and $\{a_j\}_{j \leq p+1}$, we have

$$V_{K}(\tilde{\epsilon}_{2}, \mathcal{D}) = V_{2}\left(\sqrt{2}\sigma\tilde{\epsilon}_{2}, \mathcal{B}\right) \leq K\left(\frac{\sqrt{2}\sigma\tilde{\epsilon}_{2}}{6}\right)$$

$$\leq (1+\epsilon_{0})\frac{1}{2}N(d,p)\ln\left(\frac{18\mu_{p}^{s}}{\sigma^{2}\tilde{\epsilon}_{2}^{2}}\right)$$

$$\leq (1+\epsilon_{0})\frac{1}{2}N(d,p)\ln\left(\frac{18\mathfrak{C}_{2}d^{-ps}}{\sigma^{2}C_{2}d^{-(p+1)s}}\right)$$

$$= (1+\epsilon_{0})\frac{1}{2}N(d,p)\left(\ln\frac{18\mathfrak{C}_{2}}{\sigma^{2}C_{2}} + s\ln d\right), \tag{167}$$

where we use Lemma 31 and Lemma 32 for the first line and use Lemma 17 for the third line.

Using (158) and (159), also recalling that we assume $c_1 d^{\gamma} \leq n \leq c_2 d^{\gamma}$, we have

$$\frac{V_K(\tilde{\epsilon}_2, \mathcal{D}) + n\tilde{\epsilon}_2^2 + \ln 2}{V_2(\tilde{\epsilon}_1, \mathcal{B})} \le \frac{(1 + \epsilon_0) \frac{1}{2} N(d, p) \left(\ln \frac{18\mathfrak{C}_2}{\sigma^2 C_2} + s \ln d \right) + c_2 d^{\gamma} \cdot C_2 d^{-(p+1)s} + \ln 2}{\frac{1}{2} N(d, p+1) \ln \frac{\mathfrak{C}_1}{C_1}}.$$
(168)

Recalling that Lemma 19 shows $N(d, p) \leq \mathfrak{C}_4 d^p$ and $N(d, p + 1) \geq \mathfrak{C}_3 d^{p+1}$ when $d \geq \mathfrak{C}$, the dominant terms in (168) are:

$$\frac{c_2 C_2 d^{\gamma - (p+1)s}}{\frac{1}{2} \ln \frac{\mathcal{C}_1}{C_1} N(d, p+1) \ln d}.$$
 (169)

Further noticing that $\gamma - (p+1)s \le p+1$ for any $\gamma \in (p+ps+s,(p+1)+(p+1)s]$, so we can choose C_2 small enough and only depending on $s, \sigma, \gamma, \kappa, c_1, c_2$, such that

$$\frac{V_K(\tilde{\epsilon}_2, \mathcal{D}) + n\tilde{\epsilon}_2^2 + \ln 2}{V_2(\tilde{\epsilon}_1, \mathcal{B})} \le (168) := \mathfrak{c} < 1. \tag{170}$$

Then using Lemma 33 again, we have

$$\min_{\hat{f}} \max_{\rho_{f^*} \in \mathcal{D}} \mathbb{E}_{(\boldsymbol{X}, \mathbf{y}) \sim \rho_{f^*}^{\otimes n}} \left\| \hat{f} - f^* \right\|_{L^2}^2 \ge \frac{1 - \mathfrak{c}}{4} \tilde{\epsilon}_1^2 = \frac{1 - \mathfrak{c}}{4} C_1 d^{-(p+1)s}.$$

Further recalling that $\mathcal{D} \subset \mathcal{P}$, we have

$$\min_{\hat{f}} \max_{\rho \in \mathcal{P}} \mathbb{E}_{(\boldsymbol{X}, \mathbf{y}) \sim \rho^{\otimes n}} \left\| \hat{f} - f_{\rho}^{*} \right\|_{L^{2}}^{2} \ge \min_{\hat{f}} \max_{\rho_{f^{*}} \in \mathcal{D}} \mathbb{E}_{(\boldsymbol{X}, \mathbf{y}) \sim \rho_{f^{*}}^{\otimes n}} \left\| \hat{f} - f^{*} \right\|_{L^{2}}^{2} \ge \frac{1 - \mathfrak{c}}{4} C_{1} d^{-(p+1)s}. \tag{171}$$

We finis the proof of Theorem 5 (ii).

D. Auxiliary results

The following proposition about estimating the L^2 norm with empirical norm is from Li et al. (2023b, Proposition C.9), which dates back to Caponnetto and Yao (2010).

Proposition 34 Let μ be a probability measure on \mathcal{X} , $f \in L^2(\mathcal{X}, \mu)$ and $||f||_{L^{\infty}} \leq M$. Suppose we have $\mathbf{x}_1, \ldots, \mathbf{x}_n$ sampled i.i.d. from μ . Then for $\delta \in (0, 1)$, the following holds with probability at least $1 - \delta$:

$$\frac{1}{2} \|f\|_{L^2}^2 - \frac{5M^2}{3n} \ln \frac{2}{\delta} \le \|f\|_{L^2, n}^2 \le \frac{3}{2} \|f\|_{L^2}^2 + \frac{5M^2}{3n} \ln \frac{2}{\delta}. \tag{172}$$

The following concentration inequality about self-adjoint Hilbert-Schmidt operator valued random variables is frequently used in related literature, e.g., Fischer and Steinwart (2020, Theorem 27) and Lin and Cevher (2020, Lemma 26).

Lemma 35 Let $(\mathcal{X}, \mathcal{B}, \mu)$ be a probability space, \mathcal{H} be a separable Hilbert space. Suppose that A_1, \dots, A_n are i.i.d. random variables with values in the set of self-adjoint Hilbert-Schmidt operators. If $\mathbb{E}A_i = 0$, and the operator norm $||A_i|| \leq L$ μ -a.e. $\mathbf{x} \in \mathcal{X}$, and there exists a self-adjoint positive semi-definite trace class operator V with $\mathbb{E}A_i^2 \leq V$. Then for $\delta \in (0,1)$, with probability at least $1 - \delta$, we have

$$\left\| \frac{1}{n} \sum_{i=1}^{n} A_i \right\| \le \frac{2L\beta}{3n} + \sqrt{\frac{2\|V\|\beta}{n}}, \quad \beta = \ln \frac{4\text{trV}}{\delta \|V\|}.$$

The following Bernstein inequality about vector-valued random variables is frequently used, e.g., Caponnetto and de Vito (2007, Proposition 2) and Fischer and Steinwart (2020, Theorem 26).

Lemma 36 (Bernstein inequality) Let (Ω, \mathcal{B}, P) be a probability space, H be a separable Hilbert space, and $\xi : \Omega \to H$ be a random variable with

$$\mathbb{E}\|\xi\|_H^m \le \frac{1}{2}m!\sigma^2L^{m-2},$$

for all m > 2. Then for $\delta \in (0,1)$, ξ_i are i.i.d. random variables, with probability at least $1 - \delta$, we have

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \xi_i - \mathbb{E}\xi \right\|_{H} \le 4\sqrt{2} \ln \frac{2}{\delta} \left(\frac{L}{n} + \frac{\sigma}{\sqrt{n}} \right).$$

Lemma 37 Given the definition of $\mathcal{N}_1(\lambda)$ as in (4). If condition (6) in Assumption 3 holds, we have

$$\|T_{\lambda}^{-\frac{1}{2}}k(\boldsymbol{x},\cdot)\|_{\mathcal{H}}^{2} \leq \mathcal{N}_{1}(\lambda), \quad \mu\text{-a.e. } \boldsymbol{x} \in \mathcal{X}.$$

$$(173)$$

Proof

$$\begin{split} \|T_{\lambda}^{-\frac{1}{2}}k(\boldsymbol{x},\cdot)\|_{\mathcal{H}}^2 &= \Big\|\sum_{i=1}^{\infty} (\frac{1}{\lambda_i + \lambda})^{\frac{1}{2}}\lambda_i e_i(\boldsymbol{x}) e_i(\cdot)\Big\|_{\mathcal{H}}^2 \\ &= \sum_{i=1}^{\infty} \frac{\lambda_i}{\lambda_i + \lambda} e_i^2(\boldsymbol{x}) \\ &\leq \mathcal{N}_1(\lambda), \quad \mu\text{-a.e. } \boldsymbol{x} \in \mathcal{X}. \end{split}$$

Lemma 37 has a direct corollary.

Lemma 38 Given the definition of $\mathcal{N}_1(\lambda)$ as in (4). If condition (6) in Assumption 3 holds, we have

$$||T_{\lambda}^{-\frac{1}{2}}T_{\boldsymbol{x}}T_{\lambda}^{-\frac{1}{2}}|| \leq \mathcal{N}_1(\lambda), \quad \mu\text{-a.e. } \boldsymbol{x} \in \mathcal{X}.$$

Proof Note that for any $f \in \mathcal{H}$,

$$\begin{split} T_{\lambda}^{-\frac{1}{2}}T_{\boldsymbol{x}}T_{\lambda}^{-\frac{1}{2}}f &= T_{\lambda}^{-\frac{1}{2}}K_{\boldsymbol{x}}K_{\boldsymbol{x}}^{*}T_{\lambda}^{-\frac{1}{2}}f\\ &= T_{\lambda}^{-\frac{1}{2}}K_{\boldsymbol{x}}\langle k(\boldsymbol{x},\cdot),T_{\lambda}^{-\frac{1}{2}}f\rangle_{\mathcal{H}}\\ &= T_{\lambda}^{-\frac{1}{2}}K_{\boldsymbol{x}}\langle T_{\lambda}^{-\frac{1}{2}}k(\boldsymbol{x},\cdot),f\rangle_{\mathcal{H}}\\ &= \langle T_{\lambda}^{-\frac{1}{2}}k(\boldsymbol{x},\cdot),f\rangle_{\mathcal{H}} \cdot T_{\lambda}^{-\frac{1}{2}}k(\boldsymbol{x},\cdot). \end{split}$$

So
$$||T_{\lambda}^{-\frac{1}{2}}T_{\boldsymbol{x}}T_{\lambda}^{-\frac{1}{2}}|| = \sup_{\|f\|_{\mathcal{H}}=1} ||T_{\lambda}^{-\frac{1}{2}}T_{\boldsymbol{x}}T_{\lambda}^{-\frac{1}{2}}f||_{\mathcal{H}} = \sup_{\|f\|_{\mathcal{H}}=1} \langle T_{\lambda}^{-\frac{1}{2}}k(\boldsymbol{x},\cdot), f \rangle_{\mathcal{H}} \cdot ||T_{\lambda}^{-\frac{1}{2}}k(\boldsymbol{x},\cdot)||_{\mathcal{H}} = ||T_{\lambda}^{-\frac{1}{2}}k(\boldsymbol{x},\cdot)||_{\mathcal{H}}^{2}.$$
 Using Lemma 37, we finish the proof.

We state the following three lemmas without proof, since the proofs are classical and the verification about the constants is tedious. We refer to Appendix A in Zhang et al. (2023a) and the reference therein for the proof, the definition of Lorentz space $L^{p,q}(\mathcal{X},\mu)$ and real interpolation $(\cdot,\cdot)_{\theta,q}$.

Lemma 39 Let μ be the probability distribution on \mathcal{X} . For $1 < p_1 \neq p_2 < \infty$, $1 \leq q \leq \infty$ and $0 < \theta < 1$, we have

$$(L^{p_1}(\mathcal{X},\mu),L^{p_2}(\mathcal{X},\mu))_{\theta,q} \cong L^{p_{\theta},q}(\mathcal{X},\mu), \quad \frac{1}{p_{\theta}} = \frac{1-\theta}{p_1} + \frac{\theta}{p_2},$$

where $L^{p_{\theta},q}(\mathcal{X},\mu)$ is the Lorentz space and the equivalent norm only involves absolute constants.

Lemma 40 Let μ be the probability distribution on \mathcal{X} . If $1 and <math>1 \le q_1 \le q_2 \le \infty$, we have

$$L^{p,q_1}(\mathcal{X},\mu) \hookrightarrow L^{p,q_2}(\mathcal{X},\mu),$$

and the operator norm are upper bounded by an absolute constant.

Lemma 41 Let μ be the probability distribution on \mathcal{X} . For 1 , we have

$$L^{p,p}(\mathcal{X},\mu) \cong L^p(\mathcal{X},\mu); \quad L^{p,\infty}(\mathcal{X},\mu) \cong L^{p,w}(\mathcal{X},\mu),$$

where $L^{p,w}(\mathcal{X},\mu)$ denotes the weak L^p space and the equivalent norm only involves absolute constants.

Theorem 42 (L^q -embedding property) Suppose that \mathcal{H} is the RKHS associated with a continuous, positive-definite and symmetric kernel k on a compact set $\mathcal{X} \subset \mathbb{R}^d$ and the probability distribution on \mathcal{X} is μ . Further suppose that $\sup_{x \in \mathcal{X}} |k(x,x)| \leq \kappa^2$, where κ is an absolute constant. Then for any 0 < s < 1, we have

$$[\mathcal{H}]^s \hookrightarrow L^{q_s}(\mathcal{X}, \mu), \quad \forall q_s < \frac{2}{1-s},$$
 (174)

and there exists a constant $C_{s,\kappa}$ only depending on s and κ , such that the operator norm of the embedding operator satisfies

$$\|[\mathcal{H}]^s \hookrightarrow L^{q_s}(\mathcal{X}, \mu)\| \le C_{s,\kappa}.$$
 (175)

Proof Denote $(E_0, E_1)_{\theta,q}$ as the real interpolation of two normed spaces. Steinwart and Scovel (2012, Theorem 4.6) shows that for 0 < s < 1,

$$[\mathcal{H}]^s \cong \left(L^2(\mathcal{X}, \mu), [\mathcal{H}]^1\right)_{s,2},\tag{176}$$

where the equivalent norm involves constants only depending on s.

Since $\sup_{\boldsymbol{x} \in \mathcal{X}} |k(\boldsymbol{x}, \boldsymbol{x})| \leq \kappa^2$ implies that the operator norm of embedding $I_1 : \mathcal{H} \hookrightarrow L^{\infty}$ satisfies $||I_1|| \leq \kappa^2$. Define $I_2 : (L^2, \mathcal{H})_{\theta, 2} \hookrightarrow (L^2, L^{\infty})_{\theta, 2}$ for some $\theta \in (0, 1)$, the definition of real interpolation through K-Method (see Chapter 22 in Tartar 2007) actually implies $||I_2|| \leq \max\{1, \kappa^2\}$. Then any $0 < M < \infty$, using Lemma 39, we have

$$[\mathcal{H}]^s \hookrightarrow (L^2(\mathcal{X}, \mu), L^M(\mathcal{X}, \mu))_{s,2} \cong L^{q'_s, 2}(\mathcal{X}, \mu),$$

where $\frac{1}{q'_s} = \frac{1-s}{2} + \frac{s}{M}$.

For any $q_s < \frac{2}{1-s}$, we can choose M large enough such that $q'_s > q_s$. Further, since 0 < s < 1 and thus $q'_s > q_s > 2$, using Lemma 40 and Lemma 41, we have

$$L^{q'_s,2}(\mathcal{X},\mu) \hookrightarrow L^{q'_s,q'_s}(\mathcal{X},\mu) \cong L^{q'_s}(\mathcal{X},\mu) \hookrightarrow L^{q_s}(\mathcal{X},\mu).$$

We finish the proof.

In the following, we provide some remark on Theorem 42.

Remark 43 Intuitively, their should be a constant depending on the dimension d in the operator norm of the embedding. For instance, their are extensive literature studying the dependence of the embedding constants on d in the Sobolev type inequalities (Cotsiolis and Tavoularis, 2004; Mizuguchi et al., 2016; Novak et al., 2018).

Denote $(I - \Delta)^{-\frac{r}{2}}, r > 0$, as the Bessel potential operators (see, e.g., Section 2 of Cotsiolis and Tavoularis 2004). Then the fractional Sobolev space can be defined as $H^r(\mathbb{R}^d) = \left\{ f \in L^2(\mathbb{R}^d) \mid \|(I - \Delta)^{\frac{r}{2}}f\|_{L^2} < \infty \right\}, r > 0$. It is well known that when $r > \frac{d}{2}$, $H^r(\mathbb{R}^d)$ is an RKHS with bounded kernel function. When $r < \frac{d}{2}$, denote $I_{d,r}$ as the embedding from $H^r(\mathbb{R}^d)$ to $L^{\frac{2d}{d-2r}}(\mathbb{R}^d)$. Theorem 1.1 in Cotsiolis and Tavoularis (2004) gives an upper bound of the

operator norm $||I_{d,r}||$ for any $d > 0, 0 < r < \frac{d}{2}$ (note that $||(-\Delta)^{\frac{r}{2}}f||_{L^2} \le ||(I-\Delta)^{\frac{r}{2}}f||_{L^2}$). Since for 0 < s < 1, $[H^{\frac{d}{2}}(\mathbb{R}^d)]^s \cong (L^2(\mathbb{R}^d), H^{\frac{d}{2}}(\mathbb{R}^d))_{s,2} \cong H^{\frac{ds}{2}}(\mathbb{R}^d)$), letting $r = \frac{d}{2}$, we have

$$H^r(\mathbb{R}^d) = [\mathcal{H}_d]^{\frac{2}{3}}, \text{ where } \mathcal{H}_d = H^{\frac{d}{2}}(\mathbb{R}^d) \text{ is an RKHS.}$$

(With a little abusement of notation, we consider $H^{\frac{d}{2}}(\mathbb{R}^d)$ as an RKHS). Then $I_{d,\frac{d}{2}}$ can also be interpreted as

$$I_{d,\frac{d}{2}}: [\mathcal{H}_d]^s \hookrightarrow L^{\frac{2}{1-s}}(\mathbb{R}^d), \text{ with } s = \frac{2}{3}.$$
 (177)

Detailed calculation about the constant in Theorem 1.1 in Cotsiolis and Tavoularis (2004) shows that the operator norm of $I_{d,\frac{d}{2}}$ decreases to 0, i.e.,

$$||I_{d,\frac{d}{2}}|| \to 0$$
, as $d \to \infty$. (178)

This indicates that although the embedding norm may depend on d, it can always be upper bounded by a constant. This shows the consistency with Theorem 42 in our paper.

So when will the embedding norm tend to 0 and when will it remain as a constant? We can get some inspiration from the operator norm of $I_1: \mathcal{H}_d \hookrightarrow L^{\infty}$. Recall that we assume $\sup_{\boldsymbol{x} \in \mathcal{X}} |k(\boldsymbol{x}, \boldsymbol{x})| \leq \kappa^2$ in Theorem 42, where κ is an absolute constant. This directly implies $\boldsymbol{x} \in \mathcal{X}$

 $||I_1|| \le \kappa^2$. This assumption is appropriate for some RKHSs, for instance, the inner product kernel in Assumption 4 and NTK on the sphere. For theses RKHSs, $\sup_{x \in \mathcal{X}} |k(x,x)| \le \kappa^2$ will

not change as $d \to \infty$. However, we conjecture that the bound $||I_1|| \le \kappa^2$ may be too loose for other RKHSs when $d \to \infty$.

Let us see the example of fractional Sobolev space again. For $d, r \in \mathbb{N}$ with $r > \frac{d}{2}$, denote $I_{d,r,\infty}$ as the embedding from $H^r(\mathbb{R}^d)$ to $L^{\infty}(\mathbb{R}^d)$. Theorem 11 in Novak et al. (2018) shows that $||I_{d,r,\infty}|| \to 0$ as $d \to \infty$. Note that the definition of $H^r(\mathbb{R}^d)$ in this section is actually the same as the definition in Section 4.1 of Novak et al. (2018).

References

- M. Aerni, M. Milanta, K. Donhauser, and F. Yang. Strong inductive biases provably prevent harmless interpolation. In *The Eleventh International Conference on Learning Representations*, 2022.
- S. Arora, S. S. Du, W. Hu, Z. Li, R. R. Salakhutdinov, and R. Wang. On exact computation with an infinitely wide neural net. *Advances in Neural Information Processing Systems*, 32, 2019.
- P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler. Benign overfitting in linear regression. Proceedings of the National Academy of Sciences, 117(48):30063–30070, 2020.
- F. Bauer, S. Pereverzyev, and L. Rosasco. On regularization algorithms in learning theory. Journal of complexity, 23(1):52–72, 2007.
- D. Beaglehole, M. Belkin, and P. Pandit. On the inconsistency of kernel ridgeless regression in fixed dimensions. SIAM Journal on Mathematics of Data Science, 5(4):854–872, 2023.
- A. Bietti and J. Mairal. On the inductive bias of neural tangent kernels. *Advances in Neural Information Processing Systems*, 32, 2019.
- B. Bordelon, A. Canatar, and C. Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, pages 1024–1034. PMLR, 2020.
- S. Buchholz. Kernel interpolation in Sobolev spaces is not consistent in low dimensions. In P.-L. Loh and M. Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 3410–3440. PMLR, July 2022.
- A. Caponnetto. Optimal rates for regularization operators in learning theory. Technical report, MASSACHUSETTS INST OF TECH CAMBRIDGE COMPUTER SCIENCE AND ARTIFICIAL ..., 2006.
- A. Caponnetto and E. de Vito. Optimal rates for the regularized least-squares algorithm. Foundations of Computational Mathematics, 7:331–368, 2007.
- A. Caponnetto and Y. Yao. Cross-validation based adaptation for regularization operators in learning theory. *Analysis and Applications*, 8(02):161–183, 2010.
- A. Cotsiolis and N. K. Tavoularis. Best constants for sobolev inequalities for higher order fractional derivatives. *Journal of mathematical analysis and applications*, 295(1):225–236, 2004.
- H. Cui, B. Loureiro, F. Krzakala, and L. Zdeborová. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. Advances in Neural Information Processing Systems, 34:10131–10143, 2021.

- F. Dai and Y. Xu. Approximation Theory and Harmonic Analysis on Spheres and Balls. Springer Monographs in Mathematics. Springer New York, New York, NY, 2013. ISBN 978-1-4614-6659-8 978-1-4614-6660-4. doi: 10.1007/978-1-4614-6660-4.
- K. Donhauser, M. Wu, and F. Yang. How rotational invariance of common kernels prevents generalization in high dimensions. In *International Conference on Machine Learning*, pages 2804–2814. PMLR, 2021.
- S.-R. Fischer and I. Steinwart. Sobolev norm learning rates for regularized least-squares algorithms. *Journal of Machine Learning Research*, 21:205:1–205:38, 2020.
- J. Gallier, J. Quaintance, J. Gallier, and J. Quaintance. Spherical harmonics and linear representations of lie groups. *Differential Geometry and Lie Groups: A Second Course*, pages 265–360, 2020.
- L. L. Gerfo, L. Rosasco, F. Odone, E. D. Vito, and A. Verri. Spectral algorithms for supervised learning. *Neural Computation*, 20(7):1873–1897, 2008.
- B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari. When do neural networks outperform kernel methods? *Advances in Neural Information Processing Systems*, 33: 14820–14830, 2020.
- B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2):1029 1054, 2021. doi: 10.1214/20-AOS1990. URL https://doi.org/10.1214/20-AOS1990.
- N. Ghosh, S. Mei, and B. Yu. The three stages of learning dynamics in high-dimensional kernel methods. In *International Conference on Learning Representations*, 2021.
- T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949 986, 2022. doi: 10.1214/21-AOS2133. URL https://doi.org/10.1214/21-AOS2133.
- H. Hu and Y. M. Lu. Sharp asymptotics of kernel ridge regression beyond the linear regime. arXiv preprint arXiv:2205.06798, 2022.
- A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.
- N. E. Karoui. The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1):1 50, 2010. doi: 10.1214/08-AOS648. URL https://doi.org/10.1214/08-AOS648.
- J. Lai, M. Xu, R. Chen, and Q. Lin. Generalization ability of wide neural networks on \mathbb{R} . $arXiv\ preprint\ arXiv:2302.05933$, 2023.
- J. Lee, L. Xiao, S. Schoenholz, Y. Bahri, R. Novak, J. Sohl-Dickstein, and J. Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. Advances in neural information processing systems, 32, 2019.

- Y. Li, H. Zhang, and Q. Lin. Kernel interpolation generalizes poorly. arXiv preprint arXiv:2303.15809, 2023a.
- Y. Li, H. Zhang, and Q. Lin. On the saturation effect of kernel ridge regression. In *International Conference on Learning Representations*, Feb. 2023b.
- Y. Li, H. Zhang, and Q. Lin. On the asymptotic learning curves of kernel ridge regression under power-law decay. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023c.
- T. Liang and A. Rakhlin. Just interpolate: Kernel "Ridgeless" regression can generalize. *The Annals of Statistics*, 48(3):1329 1347, 2020. doi: 10.1214/19-AOS1849. URL https://doi.org/10.1214/19-AOS1849.
- T. Liang, A. Rakhlin, and X. Zhai. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Conference on Learning Theory*, pages 2683–2711. PMLR, 2020.
- J. Lin and V. Cevher. Optimal convergence for distributed learning with stochastic gradient methods and spectral algorithms. *Journal of Machine Learning Research*, 21:147–1, 2020.
- J. Lin, A. Rudi, L. Rosasco, and V. Cevher. Optimal rates for spectral algorithms with least-squares regression over Hilbert spaces. Applied and Computational Harmonic Analysis, 48:868–890, 2018.
- F. Liu, Z. Liao, and J. Suykens. Kernel regression in high dimensions: Refined analysis beyond double descent. In *International Conference on Artificial Intelligence and Statistics*, pages 649–657. PMLR, 2021.
- W. Lu, H. Zhang, Y. Li, M. Xu, and Q. Lin. Optimal rate of kernel regression in large dimensions. arXiv preprint arXiv:2309.04268, 2023.
- S. Mei and A. Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. Communications on Pure and Applied Mathematics, 75(4):667–766, 2022.
- S. Mei, T. Misiakiewicz, and A. Montanari. Generalization error of random feature and kernel methods: Hypercontractivity and kernel matrix concentration. Applied and Computational Harmonic Analysis, 59:3–84, 2022.
- T. Misiakiewicz. Spectrum of inner-product kernel matrices in the polynomial regime and multiple descent phenomenon in kernel ridge regression. arXiv preprint arXiv:2204.10425, 2022.
- M. Mizuguchi, A. Takayasu, T. Kubo, and S. Oishi. On the embedding constant of the sobolev type inequality for fractional derivatives. *Nonlinear Theory and Its Applications*, *IEICE*, 7(3):386–394, 2016.
- V. Muthukumar, K. Vodrahalli, V. Subramanian, and A. Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1(1): 67–83, 2020. doi: 10.1109/JSAIT.2020.2984716.

- E. Novak, M. Ullrich, H. Woźniakowski, and S. Zhang. Reproducing kernels of sobolev spaces on r d and applications to embedding constants and tractability. *Analysis and Applications*, 16(05):693–715, 2018.
- A. Rakhlin and X. Zhai. Consistency of interpolation with laplace kernels is a highdimensional phenomenon. In *Conference on Learning Theory*, pages 2595–2623. PMLR, 2019.
- I. Steinwart and A. Christmann. Support vector machines. In *Information Science and Statistics*, 2008.
- I. Steinwart and C. Scovel. Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs. *Constructive Approximation*, 35(3):363–417, 2012.
- I. Steinwart, D. Hush, and C. Scovel. Optimal rates for regularized least squares regression. In COLT, pages 79–93, 2009.
- L. Tartar. An introduction to Sobolev spaces and interpolation spaces, volume 3. Springer Science & Business Media, 2007.
- A. Tsigler and P. L. Bartlett. Benign overfitting in ridge regression. *Journal of Machine Learning Research*, 24(123):1–76, 2023.
- L. Xiao, H. Hu, T. Misiakiewicz, Y. Lu, and J. Pennington. Precise learning curves and higher-order scaling limits for dot product kernel regression. In *Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS)*, 2022.
- Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. The Annals of Statistics, 27(5):1564 1599, 1999. doi: 10.1214/aos/1017939142. URL https://doi.org/10.1214/aos/1017939142.
- H. Zhang, Y. Li, and Q. Lin. On the optimality of misspecified spectral algorithms. arXiv preprint arXiv:2303.14942, 2023a.
- H. Zhang, Y. Li, W. Lu, and Q. Lin. On the optimality of misspecified kernel ridge regression. arXiv preprint arXiv:2305.07241, 2023b.