On the optimality of misspecified spectral algorithms

Haobo Zhang zhang-hb21@mails.tsinghua.edu.cn

Yicheng Li * liyc22@mails.tsinghua.edu.cn
Center for Statistical Science, Department of Industrial Engineering
Tsinghua University
Beijing, China

Qian Lin † qianlin@tsinghua.edu.cn
Center for Statistical Science, Department of Industrial Engineering
Tsinghua University
and
Beijing Academy of Artificial Intelligence
Beijing, China

arXiv:2303.14942v1 [math.ST] 27 Mar 2023

Abstract

In the misspecified spectral algorithms problem, researchers usually assume the underground true function $f_\rho^* \in [\mathcal{H}]^s$, a less-smooth interpolation space of a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ for some $s \in (0,1)$. The existing minimax optimal results require $\|f_\rho^*\|_{L^\infty} < \infty$ which implicitly requires $s > \alpha_0$ where $\alpha_0 \in (0,1)$ is the embedding index, a constant depending on $\mathcal{H}$. Whether the spectral algorithms are optimal for all $s \in (0,1)$ is an outstanding problem lasting for years. In this paper, we show that spectral algorithms are minimax optimal for any $\alpha_0 - \frac{1}{\beta} < s < 1$, where β is the eigenvalue decay rate of $\mathcal{H}$. We also give several classes of RKHSs whose embedding index satisfies $\alpha_0 = \frac{1}{\beta}$. Thus, the spectral algorithms are minimax optimal for all $s \in (0,1)$ on these RKHSs.

Keywords: kernel methods, spectral algorithms, misspecified, reproducing kernel Hilbert space, minimax optimality,

1. Introduction

Suppose that the samples $\{(x_i, y_i)\}_{i=1}^n$ are i.i.d. sampled from an unknown distribution ρ on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}$. One of the goals of non-parametric least-squares regression is to find a function $\hat{f}$ based on the n samples such that the risk

$$\mathcal{E}(\hat{f}) = \mathbb{E}_{(x,y)\sim\rho}\left[\left(\hat{f}(x) - y\right)^2\right] \tag{1}$$

is relatively small. It is well known that the conditional mean function given by $f_\rho^*(x) \coloneqq \mathbb{E}_\rho[\, y \mid x \,] = \int_{\mathcal{Y}} y \mathrm{d}\rho(y|x)$ minimizes the risk $\mathcal{E}(f)$. Therefore, we may focus on establishing the convergence rate (either in expectation or in probability) for the excess risk (L^2-norm

∗. Haobo Zhang and Yicheng Li contributed equally to this work.
†. Corresponding author

generalization error)

$$\mathbb{E}_{x\sim\mu}\left[\left(\hat{f}(x)-f_\rho^*(x)\right)^2\right], \tag{2}$$

where μ is the marginal distribution of ρ on $\mathcal{X}$.

In the non-parametric regression settings, researchers often assume that $f_\rho^*(x)$ falls into a class of functions with a certain structure and develop non-parametric methods to obtain the estimator $\hat{f}$. One of the most popular non-parametric regression methods, the kernel method, aims to estimate f_ρ^* using candidate functions from a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$, a separable Hilbert space associated with a kernel function k defined on $\mathcal{X}$, e.g., Kohler and Krzyżak (2001); Cucker and Smale (2001); Steinwart and Christmann (2008). This paper focuses on a class of kernel methods called the *spectral algorithms* to construct the estimator of f_ρ^*.

Since the minimax optimality of spectral algorithms has been proved for the attainable case $\left(f_\rho^*\in\mathcal{H}\right)$ (Caponnetto, 2006; Caponnetto and de Vito, 2007, etc.), a large body of literature has studied the convergence rate of the generalization error of misspecified spectral algorithms ($f_\rho^*\notin\mathcal{H}$) and whether the rate is optimal in the minimax sense. It turns out that the qualification of the algorithm ($\tau>0$), the eigenvalue decay rate ($\beta>1$), the source condition ($s>0$) and the embedding index ($\alpha_0<1$) of the RKHS jointly determine the convergence behaviors of the spectral algorithms (see Section 3.1 for definitions). If we only assume that f_ρ^* belongs to an interpolation space $[\mathcal{H}]^s$ of the RKHS $\mathcal{H}$ for some $s>0$, the well known information-theoretic lower bound shows that the minimax lower bound (with respect to the L^2-norm generalization error) is $n^{-\frac{s\beta}{s\beta+1}}$. The state-of-the-art result shows that when $\alpha_0<s\le 2\tau$, the upper bound of the convergence rate (with respect to the L^2-norm generalization error) is $n^{-\frac{s\beta}{s\beta+1}}$ and hence is optimal (Fischer and Steinwart 2020 for kernel ridge regression and Pillaud-Vivien et al. 2018 for gradient methods). However, when $f_\rho^*\in[\mathcal{H}]^s$ for some $0<s\le\alpha_0$, all the existing works need an additional boundedness assumption of f_ρ^* to prove the same upper bound $n^{-\frac{s\beta}{s\beta+1}}$. The boundedness assumption will result in a smaller function space, i.e., $[\mathcal{H}]^s\cap L^\infty(\mathcal{X},\mu)\subsetneq[\mathcal{H}]^s$ when $s\le\alpha_0$. Fischer and Steinwart (2020) further reveals that the minimax rate associated with the smaller function space is larger than $n^{-\frac{\alpha\beta}{\alpha\beta+1}}$ for any $\alpha>\alpha_0$. This minimax lower bound is smaller than the upper bound of the convergence rate and hence they can not prove the minimax optimality of spectral algorithms when $s\le\alpha_0$.

It has been an outstanding problem for years whether the spectral algorithms are minimax optimal for all $s\in(0,1)$, either with respect to the L^2-norm or the $[\mathcal{H}]^\gamma$-norm introduced later (Pillaud-Vivien et al., 2018; Fischer and Steinwart, 2020; Liu and Shi, 2022). To this end, this paper has three contributions.

- Using the tools from real interpolation theory, we analyze the L^q-embedding property of $[\mathcal{H}]^s$, an interpolation space of the RKHS. Specifically, assume that $\mathcal{H}$ has embedding index α_0. When $s\le\alpha_0$, Theorem 5 proves that $[\mathcal{H}]^s$ is continuously embedded into $L^q(\mathcal{X},\mu)$, for $q=\frac{2\alpha}{\alpha-s},\forall\alpha>\alpha_0$.
- Based on the L^q-embedding property of $[\mathcal{H}]^s$, the refined proof in this paper removes the boundedness assumption in previous literature and obtains the same upper bound of the convergence rate as the state-of-the-art upper bound. As a result, we prove the

minimax optimality of spectral algorithms for $\alpha_0 - \frac{1}{\beta} < s \leq 2\tau$, which can only be proved for $\alpha_0 < s \leq 2\tau$ before. We also recover the upper bound in previous literature when $0 < s \leq \alpha_0 - \frac{1}{\beta}$ (if exists) though the optimality does not hold. Note that in this paper, we present the results in terms of $[\mathcal{H}]^\gamma$-norm generalization error, where the L^2-norm (2) is a special case when $\gamma = 0$.

- We give several examples of RKHSs whose embedding index satisfies $\alpha_0 = \frac{1}{\beta}$. Besides RKHS with uniformly bounded eigenfunctions and the Sobolev RKHS (Fischer and Steinwart, 2020), we first show that RKHS with shift-invariant kernels and RKHS with dot-product kernels on the sphere satisfy that $\alpha_0 = \frac{1}{\beta}$. Therefore, for these RKHSs, this paper proves the optimality of spectral algorithms for all $0 < s \leq 2\tau$.

1.1 Related work

General spectral algorithms in the setting of kernel methods were first proposed and studied by Rosasco et al. (2005); Caponnetto (2006); Bauer et al. (2007); Gerfo et al. (2008). A large class of regularization methods are introduced collectively as spectral algorithms and are characterized through the corresponding filter functions. The qualification τ of a spectral algorithm and a prior assumption on f_ρ^* characterizing the relative smoothness (source condition s) are also introduced for the problem setting. In this setting, Bauer et al. (2007) proves the upper bound of the convergence rate with respect to the L^2-norm generalization error. Caponnetto (2006) proves the 'capacity-dependent' upper bound, i.e., considering the eigenvalue decay rate β of the RKHS, which has been adopted by most of the later researchers. Note that these works focus on the well specified case ($f_\rho^* \in \mathcal{H}$) or assume that $\mathcal{H}$ is dense in $L^2(\mathcal{X}, \mu)$. There are also other related works studying the well specified case, e.g., Blanchard and Mücke (2018); Dicker et al. (2017); Rastogi and Sampath (2017) for general spectral algorithms, Caponnetto and de Vito (2007); Smale and Zhou (2007) for kernel ridge regression and Yao et al. (2007) for gradient methods.

Since the convergence rates and the minimax optimality of spectral algorithms in the well specified case are clear, a large amount of literature studied the misspecified spectral algorithms. Among these work, Steinwart et al. (2009); Dicker et al. (2017); Pillaud-Vivien et al. (2018); Fischer and Steinwart (2020); Celisse and Wahl (2020); Li et al. (2022); Talwai and Simchi-Levi (2022) consider the L^∞-embedding property, while Dieuleveut and Bach (2016); Lin et al. (2018); Lin and Cevher (2020); Wang and Jing (2022) do not. Note that considering the L^∞-embedding property is equivalent to introducing the embedding index α_0 in this paper. It has been shown that this will lead to faster convergence rates for certain embedding indexes (see Section 6 for detailed comparison). In addition, as mentioned in Fischer and Steinwart (2020), the convergence rates with respect to the L^2-norm can be easily extended to the more general $[\mathcal{H}]^\gamma$-norm if one uses the integral operator technique. Up to now, we have introduced five indexes τ, s, β, α_0 and γ that we know as a priori to study the convergence rates of the spectral algorithms. To our knowledge, the state-of-the-art results on the convergence rates and the minimax optimality are Fischer and Steinwart (2020) for kernel ridge regression and Pillaud-Vivien et al. (2018) for gradient methods.

But the spectral algorithms in the misspecified case have not been totally solved. When f_ρ^* falls into a less-smooth interpolation space which does not imply the boundedness of functions therein, all existing works (either considering embedding index or not) require an

additional boundedness assumption, i.e., $\|f_\rho^*\|_{L^\infty(\mathcal{X},\mu)} \leq B_\infty < \infty$ to prove the desired upper bound. As discussed in the introduction, this will lead to the suboptimality in the $s \leq \alpha_0$ regime. As far as we know, the L^q-embedding property of $[\mathcal{H}]^s$ has not been discussed in related literature. This paper shows that it turns out to be a crucial property to remove the boundedness assumption and extend the minimax optimality to a broader regime.

The outline of the rest of the paper is as follows. In Section 2, we introduce basic concepts of RKHS, integral operators and spectral algorithms. In Section 3, we present our main results of the convergence rates and the minimax optimality. As examples, we further show that the embedding index satisfies $\alpha_0 = \frac{1}{\beta}$ for some commonly used RKHSs in Section 4. Note that this is the ideal case where the minimax optimality can be proved for all $0 < s \leq 2\tau$. We verify our results through experiments in Section 5 and make a comparison with previous literature in Section 6. All the proofs can be found in Section 7.

2. Preliminaries

2.1 Basic concepts

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 ρ be an unknown probability distribution on $\mathcal{X} \times \mathcal{Y}$ satisfying $\int_{\mathcal{X}\times\mathcal{Y}} y^2 \mathrm{d}\rho(x,y) < \infty$ and denote the corresponding marginal distribution on $\mathcal{X}$ as μ. We use $L^p(\mathcal{X},\mu)$ (in short L^p) to represent the L^p-spaces. Denote

$$f_\rho^*(x) := \mathbb{E}_\rho[\ y \mid x\ ] = \int_{\mathcal{Y}} y \ \mathrm{d}\rho(y|x)$$

as the conditional mean. Throughout the paper, we denote $\mathcal{H}$ as a separable RKHS on $\mathcal{X}$ with respect to a continuous kernel function k and satisfying

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

Denote the natural embedding inclusion operator as $S_k : \mathcal{H} \to L^2(\mathcal{X},\mu)$. Moreover, the 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 $x \in \mathcal{X}$, we have

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

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

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

Next, we can define two integral operators:

$$L_k := S_k S_k^* : L^2(\mathcal{X},\mu) \to L^2(\mathcal{X},\mu), \qquad T := S_k^* S_k : \mathcal{H} \to \mathcal{H}. \tag{3}$$

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

$$L_k = \sum_{i\in N} \lambda_i \langle \cdot, e_i\rangle_{L^2} e_i \quad \text{and} \quad T = \sum_{i\in N} \lambda_i \left\langle \cdot, \lambda_i^{1/2} e_i \right\rangle_{\mathcal{H}} \lambda_i^{1/2} e_i. \tag{4}$$

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

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

where the convergence is absolute and uniform.

We also 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

$$[\mathcal{H}]^s := \operatorname{Ran} L_k^{s/2} = \left\{ \sum_{i\in N} a_i \lambda_i^{s/2} e_i : (a_i)_{i\in N} \in \ell_2(N) \right\} \subseteq L^2(\mathcal{X},\mu), \tag{5}$$

equipped with the inner product

$$\langle f, g\rangle_{[\mathcal{H}]^s} = \left\langle L_k^{-\frac{s}{2}} f, L_k^{-\frac{s}{2}} g \right\rangle_{L^2}. \tag{6}$$

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.

It is worth pointing out the relation between the definition (5) and the interpolation space defined through the real method (real interpolation). For details of interpolation of Banach spaces through the real method, we refer to Sawano (2018, Chapter 4.2.2). Specifically, Steinwart and Scovel (2012, Theorem 4.6) reveals that for $0 < s < 1$,

$$[\mathcal{H}]^s \cong \left(L^2(\mathcal{X},\mu), [\mathcal{H}]^1\right)_{s,2}. \tag{7}$$

As an example, the Sobolev space $H^m(\mathcal{X})$ is an RKHS if $m > \frac{d}{2}$ and its interpolation space is still a Sobolev space given by $[H^m(\mathcal{X})]^s \cong H^{ms}(\mathcal{X}), \forall s > 0$. See Section 4.2 for detailed discussions.

2.2 Spectral algorithms

Suppose that we observed the given samples $Z = \{(x_i, y_i)\}_{i=1}^n$ and denote $X = (x_1, \cdots, x_n)$. Define the sampling operator $K_x : \mathbb{R} \to \mathcal{H},\ y \mapsto yk(x, \cdot)$ and its adjoint operator $K_x^* : \mathcal{H} \to \mathbb{R},\ f \mapsto f(x)$. Then we can define $T_x = K_x K_x^*$. Further, we define the sample covariance operator $T_X : \mathcal{H} \to \mathcal{H}$ as

$$T_X := \frac{1}{n} \sum_{i=1}^n K_{x_i} K_{x_i}^*. \tag{8}$$

Then we know that $\|T_X\| \leq \|T_X\|_1 \leq \kappa^2$, where $\|\cdot\|$ denotes the operator norm and $\|\cdot\|_1$ denotes the trace norm. Further, define the sample basis function

$$g_Z := \frac{1}{n} \sum_{i=1}^n K_{x_i} y_i \in \mathcal{H}.$$

Based on the n samples, the kernel method aims to choose a function $\hat{f} \in \mathcal{H}$ such that the risk given by (1) is small. A direct estimator is $\hat{f} \in \mathcal{H}$ that minimizing the empirical risk

$$\hat{\mathcal{E}}(f) = \frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2,$$

which leads to an equation

$$T_X \hat{f} = g_Z.$$

However, on the one hand, minimizing the empirical risk may lead to overfitting. On the other hand, the inverse of the sample covariance operator T_X does not exist in general. The spectral algorithms (Rosasco et al., 2005; Caponnetto, 2006; Bauer et al., 2007; Gerfo et al., 2008, etc.) handle these issues by introducing the regularization and generate estimators through the filter functions. Now, we first define the filter function.

Definition 1 (Filter function) *Let $\left\{\varphi_\nu : \left[0, \kappa^2\right] \to \mathbb{R}^+ \mid \nu \in \Gamma \subseteq \mathbb{R}^+\right\}$ be a class of functions and $\psi_\nu(z) = 1 - z\varphi_\nu(z)$. If φ_ν and ψ_ν satisfy:*

- $\forall \alpha \in [0, 1]$, *we have*

$$\sup_{z \in [0, \kappa^2]} z^\alpha \varphi_\nu(z) \leq E \nu^{1-\alpha}, \quad \forall \nu \in \Gamma; \tag{9}$$

- $\exists \tau \geq 1$ *s.t.* $\forall \alpha \in [0, \tau]$, *we have*

$$\sup_{z \in [0, \kappa^2]} |\psi_\nu(z)|\, z^\alpha \leq F_\tau \nu^{-\alpha}, \quad \forall \nu \in \Gamma, \tag{10}$$

where E, F_τ are absolute constants, then we call φ_ν a filter function. We refer to ν as the regularization parameter and τ as the qualification.

Given a filter function φ_ν, we can define the corresponding spectral algorithm.

Definition 2 (spectral algorithm) *Let φ_ν be a filter function index with $\nu > 0$. Given the samples Z, the spectral algorithm produces an estimator of f_ρ^* given by*

$$\hat{f}_\nu = \varphi_\nu(T_X) g_Z. \tag{11}$$

Here we list three kinds of spectral algorithms that are commonly used.

Example 1 (Kernel ridge regression) *Let the filter function φ_ν be defined as*

$$\varphi_\nu^{\mathrm{krr}}(z) = \frac{\nu}{\nu z + 1}.$$

Then the corresponding spectral algorithm is kernel ridge regression (Tikhonov regularization). The qualification $\tau = 1$ and $E = F_\tau = 1$.

Example 2 (Gradient flow) *Let the filter function φ_ν be defined as*

$$\varphi_\nu^{\mathrm{gf}}(z) = \frac{1 - e^{-\nu z}}{z}.$$

Then the corresponding spectral algorithm is gradient flow. The qualification τ could be any positive number, $E = 1$ and $F_\tau = (\tau/e)^\tau$.

Example 3 (Spectral cut-off) *Let the filter function φ_ν be defined as*

$$\varphi_\nu^{\mathrm{cut}}(z) = \begin{cases} z^{-1}, & z^{-1} \le \nu, \\ 0, & z^{-1} > \nu. \end{cases}$$

Then the corresponding spectral algorithm is Spectral cut-off (truncated singular value decomposition). The qualification τ could be any positive number and $E = F_\tau = 1$.

For other examples of spectral algorithms (e.g., iterated Tikhonov, gradient methods, Landweber iteration, etc.), we refer to Gerfo et al. (2008).

3. Main results

3.1 Assumptions

This subsection lists the standard assumptions that frequently appear in related literature.

Assumption 1 (Eigenvalue decay rate (EDR)) *Suppose that the eigenvalue decay rate (EDR) of $\mathcal{H}$ is $\beta > 1$, i.e, there are positive constants c and C such that*

$$ci^{-\beta} \le \lambda_i \le Ci^{-\beta}, \quad \forall i \in N.$$

Note that the eigenvalues λ_i and EDR are only determined by the marginal distribution μ and the RKHS $\mathcal{H}$. The polynomial eigenvalue decay rate assumption is standard in related literature and is also referred to as the capacity condition or effective dimension condition (Caponnetto, 2006; Caponnetto and de Vito, 2007, etc.).

We say that $\mathcal{H}$ has the embedding property of order $\alpha \in [\frac{1}{\beta}, 1]$, if there is a constant $0 < A < \infty$ such that

$$\|[\mathcal{H}]^\alpha \hookrightarrow L^\infty(\mathcal{X}, \mu)\| \leq A, \tag{12}$$

where $\|\cdot\|$ denotes the operator norm of the embedding.

In fact, for any $\alpha > 0$, we can define M_α as the smallest constant $A > 0$ such that

$$\sum_{i \in N} \lambda_i^\alpha e_i^2(x) \leq A^2, \quad \mu\text{-a.e. } x \in \mathcal{X}, \tag{13}$$

if there is no such constant, set $M_\alpha = \infty$. Then Fischer and Steinwart (2020, Theorem 9) shows that for $\alpha > 0$,

$$\|[\mathcal{H}]^\alpha \hookrightarrow L^\infty(\mathcal{X}, \mu)\| = M_\alpha.$$

Note that since $\sup_{x \in \mathcal{X}} k(x, x) \leq \kappa^2$, $M_\alpha \leq \kappa < \infty$ is always true for $\alpha \geq 1$. In addition, Fischer and Steinwart (2020, Lemma 10) also shows that α can not be less than $\frac{1}{\beta}$. By the inclusion relation of interpolation spaces, it is clear that if $\mathcal{H}$ has the embedding property of order α, then it has the embedding property of order α' for any $\alpha' \geq \alpha$. Thus, we may introduce the following assumption:

Assumption 2 (Embedding index) *Suppose that there exists $\alpha_0 > 0$, such that*

$$\alpha_0 = \inf \left\{ \alpha \in [\frac{1}{\beta}, 1] : \|[\mathcal{H}]^\alpha \hookrightarrow L^\infty(\mathcal{X}, \mu)\| < \infty \right\},$$

and we refer to α_0 as the embedding index of an RKHS $\mathcal{H}$.

Note that $\mathcal{H}$ has the embedding property of order α for any $\alpha > \alpha_0$. This directly implies that all the functions in $[\mathcal{H}]^\alpha$ are μ-a.e. bounded, $\alpha > \alpha_0$. However, the embedding property may not hold for $\alpha = \alpha_0$.

Assumption 3 (Source condition) *For $s > 0$, there is a constant $R > 0$ such that $f_\rho^* \in [\mathcal{H}]^s$ and*

$$\|f_\rho^*\|_{[\mathcal{H}]^s} \leq R.$$

Functions in $[\mathcal{H}]^s$ with smaller s are less smooth, which will be harder for an algorithm to estimate.

Assumption 4 (Moment of error) *The noise $\epsilon := y - f_\rho^*(x)$ satisfies that there are constants $\sigma, L > 0$ such that for any $m \geq 2$,*

$$\mathbb{E}\left(|\epsilon|^m \mid x\right) \leq \frac{1}{2} m! \sigma^2 L^{m-2}, \quad \mu\text{-a.e. } x \in \mathcal{X}.$$

This is a standard assumption to control the noise such that the tail probability decays fast (Lin and Cevher, 2020; Fischer and Steinwart, 2020). It is satisfied for, for instance, the Gaussian noise with bounded variance or sub-Gaussian noise. Some literature (e.g., Steinwart et al. 2009; Pillaud-Vivien et al. 2018; Jun et al. 2019, etc) also uses a stronger assumption $y \in [-L_0, L_0]$ which implies both Assumption 4 and the boundedness of f_ρ^*.

3.2 Convergence results

Now we are ready to state our main results. Though this paper focuses the misspecified case, i.e., $0 < s < 1$, we state the theorems including those $s \geq 1$ for completeness.

Theorem 1 (Upper bound) *Suppose that Assumption 1,2, 3 and 4 hold for $0 < s \leq 2\tau$ and $\frac{1}{\beta} \leq \alpha_0 < 1$. Let $\hat{f}_\nu$ be the estimator defined by (11). Then for $0 \leq \gamma \leq 1$ with $\gamma \leq s$:*

- *In the case of $s + \frac{1}{\beta} > \alpha_0$, by choosing $\nu \asymp n^{\frac{\beta}{s\beta+1}}$, for any fixed $\delta \in (0,1)$, when n is sufficiently large, with probability at least $1-\delta$, we have*

$$\left\| \hat{f}_\nu - f_\rho^* \right\|_{[\mathcal{H}]^\gamma}^2 \leq \left(\ln \frac{6}{\delta} \right)^2 C n^{-\frac{(s-\gamma)\beta}{s\beta+1}}, \tag{14}$$

  where C is a constant independent of n and δ.

- *In the case of $s + \frac{1}{\beta} \leq \alpha_0$, for any $\alpha > \alpha_0$, by choosing $\nu \asymp (\frac{n}{\ln^r(n)})^{\frac{1}{\alpha}}$ for some $r > 1$, for any fixed $\delta \in (0,1)$, when n is sufficiently large, with probability at least $1-\delta$, we have*

$$\left\| \hat{f}_\nu - f_\rho^* \right\|_{[\mathcal{H}]^\gamma}^2 \leq \left(\ln \frac{6}{\delta} \right)^2 C \left(\frac{n}{\ln^r(n)} \right)^{-\frac{s-\gamma}{\alpha}}, \tag{15}$$

  where C is a constant independent of n and δ.

Compared with the state-of-the-art results (Fischer and Steinwart 2020; Pillaud-Vivien et al. 2018), Theorem 1 removes the boundedness assumption $\|f_\rho^*\|_{L^\infty} \leq B_\infty < \infty$ and prove the same upper bound for general spectral algorithms. This improvement is nontrivial for $s \leq \alpha_0$, since $[\mathcal{H}]^s \cap L^\infty(\mathcal{X}, \mu) \subsetneqq [\mathcal{H}]^s$ when $s \leq \alpha_0$. As we will see in Section 7, the proof of Theorem 1 removes the boundedness assumption by analyzing the L^q-embedding property of $[\mathcal{H}]^s$. With the L^q-integrability of the functions in $[\mathcal{H}]^s$, although the true function f_ρ^* may not fall into $L^\infty(\mathcal{X}, \mu)$, the tail probability can be controlled appropriately. We present the convergence results for $[\mathcal{H}]^\gamma$-norm generalization error, where the L^2-norm (2) is a special case when $\gamma = 0$.

Now we are going to state the minimax lower bound, which is often referred to as the information-theoretic lower bound (see, e.g., Rastogi and Sampath 2017).

Theorem 2 (Lower bound) *Let μ be a probability distribution on $\mathcal{X}$ such that Assumption 1 is satisfied. Let $\mathcal{P}$ consist of all the distributions on $\mathcal{X} \times \mathcal{Y}$ satisfying 3, 4 for $s > 0$ and with marginal distribution μ. Then for $0 \leq \gamma \leq 1$ with $\gamma \leq s$, there exists a constant C, for all learning methods, for any fixed $\delta \in (0,1)$, when n is sufficiently large, there is a distribution $\rho \in \mathcal{P}$ such that, with probability at least $1-\delta$, we have*

$$\left\| \hat{f} - f_\rho^* \right\|_{[\mathcal{H}]^\gamma}^2 \geq C\delta n^{-\frac{(s-\gamma)\beta}{s\beta+1}}. \tag{16}$$

Remark 3 (Optimality) *A direct result from Theorem 1 and Theorem 2 is that for $s \in \left(\alpha_0 - \frac{1}{\beta}, 2\tau\right]$, the upper bound matches the minimax lower bound. Therefore, we prove the minimax optimality of spectral algorithms for $s \in \left(\alpha_0 - \frac{1}{\beta}, 2\tau\right]$ with respect to $[\mathcal{H}]^\gamma$-norm ($0 \leq \gamma \leq s$) generalization error.*

4. Examples: RKHS with embedding index $\alpha_0 = \frac{1}{\beta}$

We prove the minimax optimality of spectral algorithms for $\alpha_0 - \frac{1}{\beta} < s \leq 2\tau$ in the last section. Therefore the embedding index α_0 of an RKHS is crucial when analyzing the optimality of the spectral algorithms. In the best case of $\alpha_0 = \frac{1}{\beta}$, only the first situation in Theorem 1 exists and we obtain the optimality for all $0 < s \leq 2\tau$. In this section, we give several examples of RKHSs with embedding index $\alpha_0 = \frac{1}{\beta}$.

4.1 RKHS with uniformly bounded eigenfunctions

RKHS with uniformly bounded eigenfunctions, i.e., $\sup_{i\in N} \|e_i\|_{L^\infty} < \infty$, are frequently considered (Mendelson and Neeman, 2010; Steinwart et al., 2009; Pillaud-Vivien et al., 2018). Fischer and Steinwart (2020, Lemma 10) has proved that this kind of RKHS satisfies $\alpha_0 = \frac{1}{\beta}$.

4.2 Sobolev RKHS

Let us first introduce some concepts of (fractional) Sobolev space (see, e.g., Adams and Fournier 2003). In this section, we assume that $\mathcal{X} \subseteq \mathbb{R}^d$ is a bounded domain with smooth boundary and Lebesgue measure ν. Denote $L^2(\mathcal{X}) \coloneqq L^2(\mathcal{X}, \nu)$ as the corresponding L^2 space. For $m \in \mathbb{N}$, we denote the usual Sobolev space $W^{m,2}(\mathcal{X})$ by $H^m(\mathcal{X})$ and $L^2(\mathcal{X})$ by $H^0(\mathcal{X})$. Then the (fractional) Sobolev space for any real number $r > 0$ can be defined through the *real interpolation*:

$$H^r(\mathcal{X}) := \left(L^2(\mathcal{X}), H^m(\mathcal{X})\right)_{\frac{r}{m},2},$$

where $m := \min\{k \in \mathbb{N} : k > r\}$. (We refer to Appendix A for the definition of real interpolation and Sawano 2018, Chapter 4.2.2 for more details). It is well known that when $r > \frac{d}{2}$, $H^r(\mathcal{X})$ is a separable RKHS with respect to a bounded kernel and the corresponding EDR is (see, e.g., Edmunds and Triebel 1996)

$$\beta = \frac{2r}{d}.$$

Furthermore, for the interpolation space of $H^r(\mathcal{X})$ under Lebesgue measure defined by (5), (7) shows that for $s > 0$,

$$[H^r(\mathcal{X})]^s = H^{rs}(\mathcal{X}).$$

The embedding theorem of (fractional) Sobolev space (see, e.g., 7.57 of Adams 1975) shows that if $d < 2(r - j)$ for some nonnegative integer j, then

$$H^r(\mathcal{X}) \hookrightarrow C^{j,\theta}(\mathcal{X}), \quad \theta = r - j - \frac{d}{2},$$

where $C^{j,\gamma}(\mathcal{X})$ denotes the Hölder space and $\hookrightarrow$ denotes the continuous embedding. Therefore for a Sobolev RKHS $\mathcal{H} = H^r(\mathcal{X}), r > \frac{d}{2}$ and any $\alpha > \frac{1}{\beta} = \frac{d}{2r}$,

$$[H^r(\mathcal{X})]^\alpha = H^{r\alpha}(\mathcal{X}) \hookrightarrow C^{0,\theta}(\mathcal{X}) \hookrightarrow L^\infty(\mathcal{X}),$$

where $\theta > 0$. So the embedding index of a Sobolev RKHS is $\alpha_0 = \frac{1}{\beta}$.

Furthermore, if we suppose that $\mathcal{H}$ is a Sobolev RKHS, i.e., $\mathcal{H} = H^r(\mathcal{X})$ for some $r > d/2$ and the distribution ρ satisfies that the marginal distribution μ on $\mathcal{X}$ has Lebesgue density $0 < c \le p(x) \le C$ for two constants c and C. Then we also know that the embedding index is $\alpha_0 = \frac{1}{\beta}$. Note that we say that the distribution μ has Lebesgue density $0 < c \le p(x) \le C$, if μ is equivalent to the Lebesgue measure ν, i.e., $\mu \ll \nu, \nu \ll \mu$ and there exist constants $c, C > 0$ such that $c \le \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \le C$.

4.3 RKHS with shift-invariant periodic kernels

Let us consider a kernel on $\mathcal{X} = [-\pi, \pi)^d$ satisfying

$$k(x, y) = g\big((x - y) \bmod [-\pi, \pi)^d\big),$$

where we denote

$$a \bmod [-\pi, \pi) = [(a + \pi) \bmod 2\pi] - \pi \ \in [-\pi, \pi),$$

and

$$(a_1, \ldots, a_d) \bmod [-\pi, \pi)^d = \big(a_1 \bmod [-\pi, \pi), \ldots, a_d \bmod [-\pi, \pi)\big).$$

We further assume that μ is the uniform distribution on $[-\pi, \pi)^d$. Then, it is shown in Beaglehole et al. (2022) that the Fourier basis $\phi_{\boldsymbol{k}}(x) = \exp(i\langle \boldsymbol{k}, x\rangle)$, $\boldsymbol{k} \in \mathbb{Z}^d$ are eigenfunctions of the integral operator T. Since $|\phi_{\boldsymbol{k}}(x)| \le 1$, that is, the eigenfunctions are uniformly bounded, we conclude that the embedding index $\alpha_0 = \frac{1}{\beta}$. We refer to Section 7.5 for more details.

4.4 RKHS with dot-product kernels

Dot-product kernels, which satisfy $k(x, y) = f(\langle x, y\rangle)$, have also raised researchers' interest in recent years for its nice property (Smola et al., 2000; Cho and Saul, 2009; Bach, 2017; Jacot et al., 2018). Let k be a dot-product kernel on $\mathcal{X} = \mathbb{S}^d$, the unit sphere in $\mathbb{R}^{d+1}$, and $\mu = \sigma$ be the uniform measure on $\mathbb{S}^d$. Then, it is well-known that k can be decomposed as

$$k(x, y) = \sum_{n=0}^{\infty} \mu_n \sum_{l=1}^{a_n} Y_{n,l}(x) Y_{n,l}(y),$$

where $\{Y_{n,l}\}$ is a set of orthonormal basis of $L^2(\mathbb{S}^d, \sigma)$ called the spherical harmonics. If polynomial decay condition $\mu_n \asymp n^{-d\beta}$ is satisfied (which is equivalent to assume the eigenvalue decay rate is β), Proposition 21 shows that the embedding index $\alpha_0 = \frac{1}{\beta}$ for the corresponding RKHS. We refer to Section 7.6 for more details.

5. Experiments

In this section, we aim to verify through experiments that when $\alpha_0 - \frac{1}{\beta} < s < \alpha_0$, for those functions f_ρ^* in $[\mathcal{H}]^s$ but not in L^∞, the spectral algorithms can still achieve the optimal convergence rate. We show the L^2-norm convergence results for two kinds of RKHSs and the three kinds of spectral algorithms mentioned in Section 2.2.

Suppose that $\mathcal{X} = [0,1]$ and the marginal distribution μ is the uniform distribution on $[0,1]$. The first considered RKHS is $\mathcal{H} = H^1(\mathcal{X})$, the Sobolev space with smoothness 1. Section 4.2 shows that the EDR is $\beta = 2$ and embedding index is $\alpha_0 = \frac{1}{\beta}$. We construct a function in $[\mathcal{H}]^s \backslash L^\infty$ by

$$f^*(x) = \sum_{k=1}^{\infty} \frac{1}{k^{s+0.5}} \left(\sin\left(2k\pi x\right) + \cos\left(2k\pi x\right)\right), \tag{17}$$

for some $0 < s < \frac{1}{\beta} = 0.5$. We will show in Appendix C that the series in (17) converges on $(0,1)$. In addition, since $\sin 2k\pi + \cos 2k\pi \equiv 1$, we also have $f^* \notin L^\infty(\mathcal{X})$. The explicit formula of the kernel associated to $H^1(\mathcal{X})$ is given by Thomas-Agnan (1996, Corollary 2), i.e., $k(x,y) = \frac{1}{\sinh 1} \cosh\left(1 - \max(x,y)\right) \cosh\left(1 - \min(x,y)\right)$.

For the second kind of RKHS, it is well known that the following RKHS

$$\mathcal{H} = \mathcal{H}_{\min}(\mathcal{X}) := \Big\{ f : [0,1] \to \mathbb{R} \mid f \text{ is A.C., } f(0) = 0, \int_0^1 \left(f'(x)\right)^2 \, \mathrm{d}x < \infty \Big\}.$$

is associated with the kernel $k(x,y) = \min(x,y)$ (Wainwright, 2019). Further, its eigenvalues and eigenfunctions can be written as

$$\lambda_n = \left(\frac{2n-1}{2}\pi\right)^{-2}, \quad n = 1, 2, \cdots$$

and

$$e_n(x) = \sqrt{2} \sin\left(\frac{2n-1}{2}\pi x\right), \quad n = 1, 2, \cdots$$

It is easy to see that the EDR is $\beta = 2$ and the eigenfunctions are uniformly bounded. Section 4.1 shows that the embedding index is $\alpha_0 = \frac{1}{\beta}$. We construct a function in $[\mathcal{H}]^s \backslash L^\infty$ by

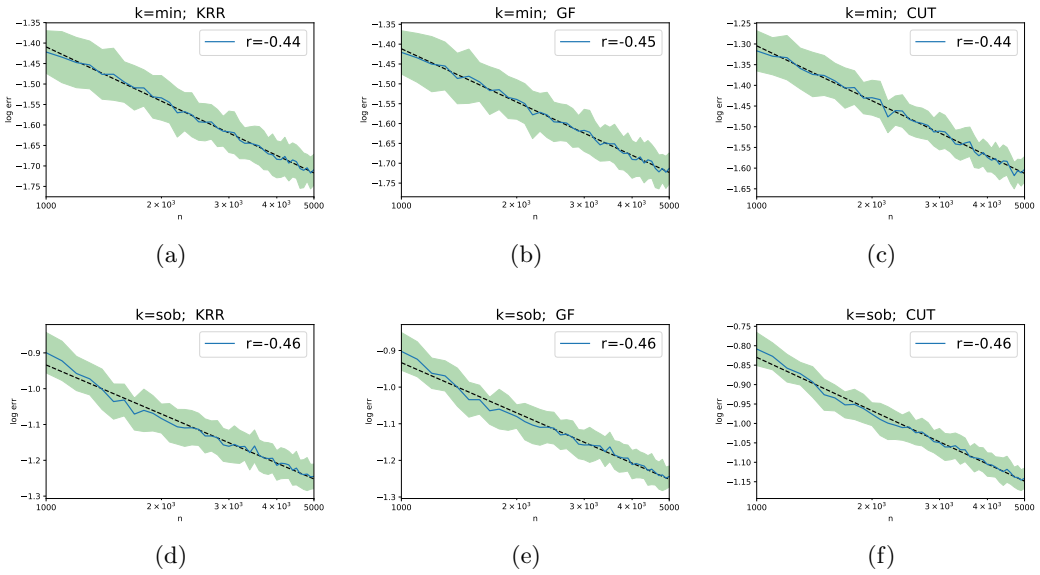
$$f^*(x) = \sum_{k=1}^{\infty} \frac{1}{k^{s+0.5}} e_{2k-1}(x), \tag{18}$$

for some $0 < s < \frac{1}{\beta} = 0.5$. We will show in Appendix C that the series in (18) converges on $(0,1)$. Since $e_{2k-1}(1) \equiv 1$, we also have $f^* \notin L^\infty(\mathcal{X})$.

We consider the following data generation procedure:

$$y = f^*(x) + \epsilon,$$

where f^* is numerically approximated by the first 3000 terms in (17) or (18) with $s = 0.4$, $x \sim \mathcal{U}[0,1]$ and $\epsilon \sim \mathcal{N}(0,1)$. Three kinds of spectral algorithms (kernel ridge regression, gradient flow and spectral cut-off) are used to construct estimators $\hat{f}$ for each RKHS, where we choose the regularization parameter as $\nu = cn^{\frac{\beta}{s\beta+1}} = cn^{\frac{10}{9}}$ for a fixed c. The sample size n is chosen from 1000 to 5000, with intervals of 100. We numerically compute the generalization error $\|\hat{f} - f^*\|_{L^2}$ by Simpson's formula with $N \gg n$ testing points. For each n, we repeat the experiments 50 times and present the average generalization error as well as the region within one standard deviation. To visualize the convergence rate r, we perform



(a) (b) (c)

(d) (e) (f)

Figure 1: Error decay curves of two kinds of RKHSs and three kinds of spectral algorithms with the best choice of c. Both axes are are scaled logarithmically. The curves show the average generalization errors over 50 trials; the regions within one standard deviation are shown in green. The dashed black lines are computed using logarithmic least-squares and the slopes represent the convergence rates r. Figures in the first row correspond to the Sobolev RKHS $\mathcal{H} = H^1(\mathcal{X})$ and the second correspond to the $\mathcal{H} = \mathcal{H}_{\min}(\mathcal{X})$.

logarithmic least-squares $\log \text{err} = r \log n + b$ to fit the generalization error with respect to the sample size and display the value of r.

We try different values of c, Figure 1 presents the convergence curves under the best choice of c. For each setting, it can be concluded that the convergence rates of the L^2-norm generalization errors of spectral algorithms are indeed approximately equal to $n^{-\frac{s\beta}{s\beta+1}} = n^{-\frac{4}{9}}$, without the boundedness assumption of the true function f^*. We refer to Appendix C for more details on the experiments.

6. Discussion

In this section, we compare this paper's convergence rates and minimax optimality with the results in previous literature. Ignoring the log-term and the constants, Theorem 1 gives the upper bound of the convergence rates of spectral algorithms (with high probability)

$$\left\| \hat{f}_\nu - f_\rho^* \right\|_{[\mathcal{H}]^\gamma}^2 \leq \begin{cases} n^{-\frac{(s-\gamma)\beta}{s\beta+1}}, & \alpha_0 - \frac{1}{\beta} < s \leq 2\tau, \\ n^{-\frac{s-\gamma}{\alpha_0+\epsilon}}, & 0 < s \leq \alpha_0 - \frac{1}{\beta}, \quad \forall \epsilon > 0. \end{cases} \tag{19}$$

This $[\mathcal{H}]^\gamma$-norm upper bound depends on τ, β, s and α_0, among which β, α_0 characterize the information of the RKHS; s characterizes the relative 'smoothness' of the true function; and τ characterizes the spectral algorithm. To our knowledge, this is the most general setting among related literature and will give the most refined analysis. In the well-specified case ($1 \leq s \leq 2\tau$ or $f_\rho^* \in \mathcal{H}$), we recover the well-known minimax optimal rates from a lot of literature (Caponnetto and de Vito, 2007; Caponnetto, 2006; Dicker et al., 2017; Blanchard and Mücke, 2018; Lin et al., 2018; Fischer and Steinwart, 2020, etc.) (for either general spectral algorithms or a specific kind).

The improvement in the misspecified case ($0 < s < 1$ or $f_\rho^* \notin \mathcal{H}$) of this paper is partly due to the advantage of considering the embedding index α_0 of the RKHS. The best upper bound without considering the embedding index is (see, e.g., Dieuleveut and Bach 2016; Lin et al. 2018; Lin and Cevher 2020)

$$\left\| \hat{f}_\nu - f_\rho^* \right\|_{[\mathcal{H}]^\gamma}^2 \leq \begin{cases} n^{-\frac{(s-\gamma)\beta}{s\beta+1}}, & 1 - \frac{1}{\beta} < s \leq 2\tau, \\ n^{-(s-\gamma)}, & 0 < s \leq 1 - \frac{1}{\beta}. \end{cases} \tag{20}$$

This rate coincides with our upper bound (19) if the embedding index $\alpha_0 = 1$. For those RKHSs with $\alpha_0 < 1$, (19) gives refined upper bound for all $0 < s \leq 2\tau$. As shown in Section 4, this is the case for many kinds of RKHSs. This is also why we assume $\alpha_0 \in (0, 1)$ throughout our paper.

Compared with the line of work which considers the embedding index (Steinwart and Christmann, 2008; Pillaud-Vivien et al., 2018; Fischer and Steinwart, 2020, etc.), this paper removes the boundedness assumption, i.e., $\|f_\rho^*\|_{L^\infty(\mathcal{X},\mu)} \leq B_\infty < \infty$. The upper bound in these works is the same as (19). But due to the boundedness assumption, Fischer and Steinwart (2020) reveals that the minimax lower rate associated to the smaller function space $[\mathcal{H}]^s \cap L^\infty(\mathcal{X}, \mu)$ is larger than

$$n^{-\frac{\max(s,\alpha)\beta}{\max(s,\alpha)\beta+1}}, \quad \forall \alpha > \alpha_0.$$

Therefore, they only prove the minimax optimality in the regime

$$\alpha_0 < s \leq 2\tau.$$

Combining Theorem 1 and Theorem 2, this paper extends the minimax optimality of the spectral algorithms to the regime

$$\alpha_0 - \frac{1}{\beta} < s \leq 2\tau.$$

This improvement is mainly due to the L^q-embedding property of the interpolation space $[\mathcal{H}]^s$ proved in Theorem 5 and a truncation method in the proof. Note that only the L^∞-embedding property has been considered before this paper. This new regime of minimax optimality means a lot. Since we have proved that the embedding index α_0 equals $\frac{1}{\beta}$ for many kinds of RKHSs, the optimality in the misspecified case is well understood for these RKHSs.

We believe that the new technical tools in this paper can be used for more related topics. For instance, some literature considers the general source condition, i.e.,

$$f_\rho^* = \phi(L_k) g_0, \quad \text{with} \ \ \|g_0\|_{L^2} \leq R,$$

where $\phi : [0, \kappa^2] \to \mathbb{R}^+$ is a non-decreasing index function such that $\phi(0) = 0$ and $\phi(\kappa^2) < \infty$ (Bauer et al., 2007; Rastogi and Sampath, 2017; Lin et al., 2018; Talwai and Simchi-Levi, 2022). The source condition in Assumption 3 corresponds to a special choice of $\phi(x) = x^{\frac{s}{2}}$, which is often referred to as the Hölder source condition. Another interesting topic is the distributed version of spectral algorithms (Zhang et al., 2013; Lin et al., 2016; Guo et al., 2017; Mücke and Blanchard, 2018; Lin and Cevher, 2020). It aims to reduce the computation complexity of the original spectral algorithms while maintaining the estimation efficiency. It would be interesting to apply this paper's tools and try refining the convergence rates or optimality in these scenarios.

In addition, we also notice a line of work which studies the learning curves of kernel ridge regression (Spigler et al., 2020; Bordelon et al., 2020; Cui et al., 2021) and crossovers between different noise levels. At present, their results all rely on a Gaussian design assumption (or some variation), which is a very strong assumption. We believe that studying the misspecified case in our paper is a crucial step to remove the Gaussian design assumption and draw complete conclusions about the learning curves of kernel ridge regression (or further, general spectral algorithms).

The eigenvalue decay rate (also known as the capacity condition or effective dimension condition) and source condition are mentioned in almost all related literature studying the convergence behaviors of kernel methods but are denoted as various kinds of notations. At the end of this section, we list a dictionary of nations in related literature. Recall that in this paper we denote the eigenvalue decay rate as β and denote the source condition as s. Table 1 summarizes the notations used in some of the references.

β	s	Reference
$1/p$	β	Steinwart et al. (2009); Fischer and Steinwart (2020); Li et al. (2022)
$1/\gamma$	2ζ	Lin et al. (2018); Lin and Cevher (2020)
b	c	Caponnetto (2006); Caponnetto and de Vito (2007)
$-$	$2r$	Bauer et al. (2007); Smale and Zhou (2007); Gerfo et al. (2008)
2ν	$\zeta + 1$	Dicker et al. (2017)
b	$2r + 1$	Rastogi and Sampath (2017); Blanchard and Mücke (2018)
$1/b$	2β	Jun et al. (2019)
α	$2r$	Dieuleveut and Bach (2016); Pillaud-Vivien et al. (2018) Celisse and Wahl (2020)

Table 1: A dictionary of notations in related literature.

7. Proofs

7.1 L^q-embedding property of the interpolation space

Before introducing the L^q-embedding property of the interpolation space $[\mathcal{H}]^s$, we first prove the following lemma, which characterizes the real interpolation between two L^p spaces with

Lorentz space $L^{p,q}(\mathcal{X},\mu)$. We refer to Appendix A for details of *real interpolation* and *Lorentz spaces*.

Lemma 4 *For $1 < p_1 \neq p_2 < \infty$, $1 \leq q \leq \infty$ and $0 < \theta < 1$, we have*

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

where $L^{p_\theta,q}(\mathcal{X},\mu)$ is the Lorentz space.

Proof Denote $L^p(\mathcal{X},\mu), L^{p,q}(\mathcal{X},\mu)$ as $L^p, L^{p,q}$ for brevity. Using Lemma 29, we know that $L^{p_i} \cong L^{p_i,p_i} = \left(L^1, L^\infty\right)_{\frac{1}{p_i'},p_i}$, where $\frac{1}{p_i'} + \frac{1}{p_i} = 1, i = 1,2$. Since $1 < p_i < \infty, i = 1,2$, Lemma 24 implies that

$$\left(L^1, L^\infty\right)_{\frac{1}{p_1'},1} \subset L^{p_1} \subset \left(L^1, L^\infty\right)_{\frac{1}{p_1'},\infty};$$
$$\left(L^1, L^\infty\right)_{\frac{1}{p_2'},1} \subset L^{p_2} \subset \left(L^1, L^\infty\right)_{\frac{1}{p_2'},\infty}.$$

Using the Reiteration theorem (Theorem 26), we have

$$(L^{p_1}, L^{p_2})_{\theta,q} = \left(L^1, L^\infty\right)_{\eta,q}, \tag{21}$$

where $\eta = \frac{1-\theta}{p_1'} + \frac{\theta}{p_2'}$. Simple calculations show that

$$1 - \eta = \frac{1-\theta}{p_1} + \frac{\theta}{p_2} := \frac{1}{p_\theta}.$$

So by the definition of Lorentz space, we have

$$\left(L^1, L^\infty\right)_{\eta,q} = L^{\frac{1}{1-\eta},q} = L^{p_\theta,q}.$$

Together with (21), we finish the proof. ■

Based on Lemma 4, the following theorem gives the L^q-embedding property of the interpolation space of an RKHS $\mathcal{H}$, which is crucial for proving the upper bound.

Theorem 5 (L^q-embedding property) *Suppose that the RKHS $\mathcal{H}$ has embedding index α_0, then for any $0 < s \leq \alpha_0$, we have*

$$[\mathcal{H}]^s \hookrightarrow L^{q_s}(\mathcal{X},\mu), \quad q_s = \frac{2\alpha}{\alpha - s}, \quad \forall \alpha > \alpha_0.$$

where $\hookrightarrow$ denotes the continuous embedding.

Proof Since the embedding index is α_0, we know that $[\mathcal{H}]^\alpha \hookrightarrow L^\infty(\mathcal{X},\mu), \forall \alpha > \alpha_0$. In addition, (7) shows that

$$[\mathcal{H}]^s = \left[\, [\mathcal{H}]^\alpha \,\right]^{\frac{s}{\alpha}} \cong \left(L^2(\mathcal{X},\mu), [\mathcal{H}]^\alpha\right)_{\frac{s}{\alpha},2}.$$

So using Lemma 4, we have

$$[\mathcal{H}]^s \hookrightarrow \big(L^2(\mathcal{X},\mu), L^\infty(\mathcal{X},\mu)\big)_{\frac{s}{\alpha},2} \cong L^{p_s,2}(\mathcal{X},\mu),$$

where $\frac{1}{p_s} = \frac{1-\frac{s}{\alpha}}{2} + \frac{\frac{s}{\alpha}}{\infty} = \frac{\alpha - s}{2\alpha}$. Further, since $0 < s \le \alpha_0 < \alpha$ thus $p_s > 2$, using Lemma 24 and Lemma 29, we have

$$L^{p_s,2}(\mathcal{X},\mu) \hookrightarrow L^{p_s,p_s}(\mathcal{X},\mu) \cong L^{p_s}(\mathcal{X},\mu).$$

We finish the proof. ■

7.2 Some bounds

Throughout the proof, we denote

$$T_\nu = T + \nu^{-1}; \quad T_{X\nu} = T_X + \nu^{-1},$$

where ν 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 $\mathrm{tr}A$ 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 use $a_n \asymp b_n$ to denote that there exist constants c and C such that $ca_n \le b_n \le Ca_n, \forall n = 1,2,\cdots$; use $a_n \lesssim b_n$ to denote that there exists an constant C such that $a_n \le Cb_n, \forall n = 1,2,\cdots$

In addition, denote the effective dimension as

$$\mathcal{N}(\nu) = \mathrm{tr}\big(T(T+\nu^{-1})^{-1}\big) = \sum_{i\in N} \frac{\lambda_i}{\lambda_i + \nu^{-1}}.$$

Since the EDR of $\mathcal{H}$ is β, Lemma 31 shows that $\mathcal{N}(\nu) \asymp \nu^{\frac{1}{\beta}}$.

Recall that we have define the sample basis function g_Z and the spectral algorithm $\hat{f}_\nu$ in Section 2.2. We also need the following notations: define the expectation of g_Z as

$$g = \mathbb{E}g_Z = \int_{\mathcal{X}} k(x,\cdot) f_\rho^*(x) d\mu(x) = S_k^* f_\rho^* \in \mathcal{H},$$

and

$$f_\nu = \varphi_\nu(T)g = \varphi_\nu(T) S_k^* f_\rho^*.$$

The following theorem bounds the $[\mathcal{H}]^\gamma$-norm of $f_\nu - f_\rho^*$ when $0 \le \gamma \le s$.

Theorem 6 *Suppose that Assumption 3 holds for $0 < s \le 2\tau$. Then for any $\nu > 0$ and $0 \le \gamma \le s$, we have*

$$\big\|f_\nu - f_\rho^*\big\|_{[\mathcal{H}]^\gamma} \le F_\tau R \nu^{-\frac{s-\gamma}{2}}. \tag{22}$$

Proof Suppose that $f_\rho^* = L_k^{\frac{s}{2}} g_0$ for some $g_0 \in L^2$. Note that

$$\begin{aligned}
\left\|f_\nu - f_\rho^*\right\|_{[\mathcal{H}]^\gamma} &= \left\| L_k^{-\frac{\gamma}{2}} \left(S_k \varphi_\nu(T) S_k^* f_\rho^* - f_\rho^*\right) \right\|_{L^2} \\
&= \left\| L_k^{-\frac{\gamma}{2}} \left(\varphi_\nu(L_k) L_k - I\right) L_k^{\frac{s}{2}} g_0 \right\|_{L^2} \\
&\leq \left\| L_k^{\frac{s-\gamma}{2}} \psi_\nu(L_k) \right\| R \\
&\leq F_\tau R \nu^{-\frac{s-\gamma}{2}},
\end{aligned}$$

where we use the property of the filter function (10) and $\|g_0\|_{L^2} = \|f_\rho^*\|_{[\mathcal{H}]^s} \leq R$ for the last inequality. ■

The following lemma bounds the L^∞-norm of f_ν when $s \leq \alpha_0$.

Lemma 7 *Suppose that Assumption 1, 2 and 3 hold for $0 < s \leq \alpha_0$ and $\frac{1}{\beta} \leq \alpha_0 < 1$. Then for any $\nu > 0$ and any $\alpha_0 < \alpha \leq 1$, we have*

$$\|f_\nu\|_{L^\infty} \leq M_\alpha E R \nu^{\frac{\alpha-s}{2}}. \tag{23}$$

Proof Since $s \leq \alpha_0$ and $\alpha > \alpha_0$, we have

$$\begin{aligned}
\|f_\nu\|_{[\mathcal{H}]^\alpha} &= \left\| L_k^{-\frac{\alpha}{2}} S_k \varphi_\nu(T) S_k^* f_\rho^* \right\|_{[\mathcal{H}]^\alpha} \\
&= \left\| L_k^{-\frac{\alpha}{2}} \varphi_\nu(L_k) L_k L_k^{\frac{s}{2}} g_0 \right\|_{L^2} \\
&= \left\| L_k^{1-\frac{\alpha-s}{2}} \varphi_\nu(L_k) g_0 \right\|_{L^2} \\
&= \left\| L_k^{1-\frac{\alpha-s}{2}} \varphi_\nu(L_k) \right\| \|g_0\|_{L^2} \\
&\leq E R \nu^{\frac{\alpha-s}{2}},
\end{aligned}$$

where we use the property of the filter function (9) for the last inequality. Further, using $\|[\mathcal{H}]^\alpha \hookrightarrow L^\infty(\mathcal{X}, \mu)\| = M_\alpha$ by Assumption 2, we have $\|f_\nu\|_{L^\infty} \leq M_\alpha \|f_\nu\|_{[\mathcal{H}]^\alpha} \leq M_\alpha E R \nu^{-\frac{\alpha-s}{2}}$. ■

The following lemma will be frequently used in our proof.

Lemma 8 *Suppose that the RKHS $\mathcal{H}$ has embedding index α_0. For any $\alpha_0 < \alpha \leq 1$, we have*

$$\|T_\nu^{-\frac{1}{2}} k(x, \cdot)\|_{\mathcal{H}}^2 \leq M_\alpha^2 \nu^\alpha, \quad \mu\text{-a.e. } x \in \mathcal{X}. \tag{24}$$

Proof Recalling the definition of the embedding index, for any $\alpha_0 < \alpha \leq 1$,

$$\sum_{i \in N} \lambda_i^\alpha e_i^2(x) \leq M_\alpha, \quad \mu\text{-a.e. } x \in \mathcal{X}.$$

So, we have

$$\begin{aligned}\|T_\nu^{-\frac{1}{2}}k(x,\cdot)\|_{\mathcal{H}}^2 &= \Big\|\sum_{i\in N}(\frac{1}{\lambda_i+\nu^{-1}})^{\frac{1}{2}}\lambda_i e_i(x)e_i(\cdot)\Big\|_{\mathcal{H}}^2\\ &= \sum_{i\in N}\frac{\lambda_i}{\lambda_i+\nu^{-1}}e_i^2(x)\\ &= \big[\sum_{i\in N}\lambda_i^\alpha e_i^2(x)\big]\sup_{i\in N}\frac{\lambda_i^{1-\alpha}}{\lambda_i+\nu^{-1}}\\ &\le M_\alpha^2\nu^\alpha, \quad \mu\text{-a.e. } x\in\mathcal{X}.\end{aligned}$$

where we use Lemma 30 for the last inequality and we finish the proof. ■

Lemma 8 has a direct corollary.

Lemma 9 *Suppose that the RKHS $\mathcal{H}$ has embedding index α_0. For any $\alpha_0<\alpha\le 1$, we have*

$$\|T_\nu^{-\frac{1}{2}}T_xT_\nu^{-\frac{1}{2}}\|\le M_\alpha^2\nu^\alpha, \quad \mu\text{-a.e. } x\in\mathcal{X}.$$

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

$$\begin{aligned}T_\nu^{-\frac{1}{2}}T_xT_\nu^{-\frac{1}{2}}f &= T_\nu^{-\frac{1}{2}}K_xK_x^*T_\nu^{-\frac{1}{2}}f\\ &= T_\nu^{-\frac{1}{2}}K_x\langle k(x,\cdot),T_\nu^{-\frac{1}{2}}f\rangle_{\mathcal{H}}\\ &= T_\nu^{-\frac{1}{2}}K_x\langle T_\nu^{-\frac{1}{2}}k(x,\cdot),f\rangle_{\mathcal{H}}\\ &= \langle T_\nu^{-\frac{1}{2}}k(x,\cdot),f\rangle_{\mathcal{H}}\cdot T_\nu^{-\frac{1}{2}}k(x,\cdot).\end{aligned}$$

So $\|T_\nu^{-\frac{1}{2}}T_xT_\nu^{-\frac{1}{2}}\| = \sup_{\|f\|_{\mathcal{H}}=1}\|T_\nu^{-\frac{1}{2}}T_xT_\nu^{-\frac{1}{2}}f\|_{\mathcal{H}} = \sup_{\|f\|_{\mathcal{H}}=1}\langle T_\nu^{-\frac{1}{2}}k(x,\cdot),f\rangle_{\mathcal{H}}\cdot\|T_\nu^{-\frac{1}{2}}k(x,\cdot)\|_{\mathcal{H}} = \|T_\nu^{-\frac{1}{2}}k(x,\cdot)\|_{\mathcal{H}}^2$. Using Lemma 8, we finish the proof. ■

The following lemma is a corollary of Lemma 33, which is also used in Lin et al. (2018, Lemma 5.5) and Smale and Zhou (2007).

Lemma 10 *Let $0<\delta<\frac{1}{2}$, it holds with probability at least $1-\delta$*

$$\|T_X-T\|\le\|T_X-T\|_2\le\frac{8\sqrt{2}\kappa^2}{\sqrt{n}}\ln\frac{2}{\delta},$$

where $\|\cdot\|$ denotes the operator norm and $\|\cdot\|_2$ denotes the Hilbert-Schmidt norm.

Proof Define $\xi(x)=T_x$, then we have

$$T_X-T=\frac{1}{n}\sum_{i=1}^n\xi(x_i)-\mathbb{E}\xi(x).$$

Since $\sup_{x\in\mathcal{X}} k(x,x) \le \kappa^2$, the Hilbert-Schmidt norm of $\xi(x)$ satisfies that

$$\|\xi(x)\|_2 \le \kappa^2, \quad \forall x \in \mathcal{X}.$$

Applying Lemma 33 with $L = \sigma = \kappa^2$, with probability at least $1-\delta$, we have

$$\|T_X - T\|_2 \le 4\sqrt{2}\ln\frac{2}{\delta}\left(\frac{\kappa^2}{n} + \frac{\kappa^2}{\sqrt{n}}\right) \le \frac{8\sqrt{2}\kappa^2}{\sqrt{n}}\ln\frac{2}{\delta}.$$

The first inequality follows from the fact that $\|T_X - T\| \le \|T_X - T\|_2$. ■

7.3 Upper bound

Lemma 11 *Suppose that the RKHS $\mathcal{H}$ has embedding index α_0. Then for any $\alpha_0 < \alpha \le 1$ and all $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\|T_\nu^{-\frac{1}{2}}(T - T_X)T_\nu^{-\frac{1}{2}}\| \le \frac{4M_\alpha^2\nu^\alpha}{3n}B + \sqrt{\frac{2M_\alpha^2\nu^\alpha}{n}}B,$$

where

$$B = \ln\frac{4\mathcal{N}(\nu)(\|T\| + \nu^{-1})}{\delta\|T\|}.$$

Proof Denote $A_i = T_\nu^{-\frac{1}{2}}(T - T_{x_i})T_\nu^{-\frac{1}{2}}$, using Lemma 9, we have

$$\|A_i\| = \|T_\nu^{-\frac{1}{2}}TT_\nu^{-\frac{1}{2}}\| + \|T_\nu^{-\frac{1}{2}}T_{x_i}T_\nu^{-\frac{1}{2}}\| \le 2M_\alpha^2\nu^\alpha, \quad \mu\text{-a.e. } x \in \mathcal{X}.$$

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

$$\mathbb{E}A_i^2 \preceq \mathbb{E}\left[T_\nu^{-\frac{1}{2}}T_{x_i}T_\nu^{-\frac{1}{2}}\right]^2.$$

In addition, Lemma 9 shows that $0 \preceq T_\nu^{-\frac{1}{2}}T_{x_i}T_\nu^{-\frac{1}{2}} \preceq M_\alpha^2\lambda^{-\alpha}, \mu$-a.e. $x \in \mathcal{X}$. So we have

$$\mathbb{E}A_i^2 \preceq \mathbb{E}\left[T_\nu^{-\frac{1}{2}}T_{x_i}T_\nu^{-\frac{1}{2}}\right]^2 \preceq \mathbb{E}\left[M_\alpha^2\nu^{-\alpha}\cdot T_\nu^{-\frac{1}{2}}T_{x_i}T_\nu^{-\frac{1}{2}}\right] = M_\alpha^2\nu^{-\alpha}T_\nu^{-1}T,$$

Define an operator $V := M_\alpha^2\nu^\alpha T_\nu^{-1}T$, we have

$$\begin{aligned}
&\|V\| = M_\alpha^2\nu^\alpha\frac{\lambda_1}{\lambda_1 + \nu^{-1}} = M_\alpha^2\nu^\alpha\frac{\|T\|}{\|T\| + \nu^{-1}} \le M_\alpha^2\nu^\alpha;\\
&\mathrm{tr}V = M_\alpha^2\nu^\alpha\mathcal{N}(\nu);\\
&\frac{\mathrm{tr}V}{\|V\|} = \frac{\mathcal{N}(\nu)(\|T\| + \nu^{-1})}{\|T\|}.
\end{aligned}$$

Use Lemma 32 to A_i, V and we finish the proof. ■

Lemma 12 *Suppose that the RKHS $\mathcal{H}$ has embedding index α_0. For any $\alpha_0 < \alpha \le 1$, if ν and n satisfy that*

$$\frac{M_\alpha^2 \nu^\alpha}{n} \ln \frac{4\kappa^2 \mathcal{N}(\nu)\left(\|T\| + \nu^{-1}\right)}{\delta \|T\|} \le \frac{1}{8}, \tag{25}$$

then for all $\delta \in (0,1)$, with probability at least $1-\delta$, we have

$$\left\| T_\nu^{-\frac{1}{2}} T_{X\nu}^{\frac{1}{2}} \right\|^2 \le 2, \quad \left\| T_\nu^{\frac{1}{2}} T_{X\nu}^{-\frac{1}{2}} \right\|^2 \le 3.$$

Proof Define

$$u = \frac{M_\alpha^2 \nu^\alpha}{n} \ln \frac{4\kappa^2 \mathcal{N}(\nu)\left(\|T\| + \nu^{-1}\right)}{\delta \|T\|} \le \frac{1}{8}.$$

Using Lemma 11, with probability at least $1-\delta$, we have

$$a := \|T_\nu^{-\frac{1}{2}}(T - T_X) T_\nu^{-\frac{1}{2}}\| \le \frac{4}{3}u + \sqrt{2u} \le \frac{2}{3}.$$

So we have

$$\begin{aligned}
\left\| T_\nu^{-\frac{1}{2}} T_{X\nu}^{\frac{1}{2}} \right\|^2 &= \left\| T_\nu^{-\frac{1}{2}} T_{X\nu} T_\nu^{-\frac{1}{2}} \right\| = \left\| T_\nu^{-\frac{1}{2}} \left(T_X + \nu^{-1}\right) T_\nu^{-\frac{1}{2}} \right\| \\
&= \left\| T_\nu^{-\frac{1}{2}} \left(T_X - T + T + \nu^{-1}\right) T_\nu^{-\frac{1}{2}} \right\| \\
&= \left\| T_\nu^{-\frac{1}{2}} \left(T_X - T\right) T_\nu^{-\frac{1}{2}} + I \right\| \\
&\le a + 1 \le 2;
\end{aligned}$$

and

$$\begin{aligned}
\left\| T_\nu^{\frac{1}{2}} T_{X\nu}^{-\frac{1}{2}} \right\|^2 &= \left\| T_\nu^{\frac{1}{2}} T_{X\nu}^{-1} T_\nu^{\frac{1}{2}} \right\| = \left\| \left(T_\nu^{-\frac{1}{2}} T_{X\nu} T_\nu^{-\frac{1}{2}} \right)^{-1} \right\| \\
&= \left\| \left(I - T_\nu^{-\frac{1}{2}} (T_X - T) T_\nu^{-\frac{1}{2}} \right)^{-1} \right\| \\
&\le \sum_{k=0}^{\infty} \left\| T_\nu^{-\frac{1}{2}} (T_X - T) T_\nu^{-\frac{1}{2}} \right\|^k \\
&\le \sum_{k=0}^{\infty} \left(\frac{2}{3} \right)^k \le 3.
\end{aligned}$$

■

The following theorem is an application of the classical Bernstein inequality but considering a truncation version of f_ρ^*, which will bring refined analysis when handling those $f_\rho^* \notin L^\infty$.

Theorem 13 *Suppose that Assumption 1, 2, 3 and 4 hold for* $0 < s \leq 2\tau$ *and* $\frac{1}{\beta} \leq \alpha_0 < 1$. *Denote* $\xi_i = \xi(x_i, y_i) = T_\nu^{-\frac{1}{2}}(K_{x_i} y_i - T_{x_i} f_\nu)$ *and* $\Omega_0 = \{x \in \Omega : |f_\rho^*(x)| \leq t\}$. *Then for any* $\alpha_0 < \alpha \leq 1$ *and all* $\delta \in (0,1)$, *with probability at least* $1-\delta$, *we have*

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \xi_i I_{x_i \in \Omega_0} - \mathbb{E} \xi_x I_{x \in \Omega_0} \right\|_{\mathcal{H}} \leq \ln \frac{2}{\delta} \left(\frac{C_1 \nu^{\frac{\alpha}{2}}}{n} \cdot \tilde{M} + \frac{C_2 \mathcal{N}^{\frac{1}{2}}(\nu)}{\sqrt{n}} + \frac{C_3 \nu^{\frac{\alpha - s}{2}}}{\sqrt{n}} \right),$$

where $\tilde{M} = M_\alpha (E + F_\tau) R \nu^{\frac{\alpha - s}{2}} + t + L$ *and* L *is the constant in (4).* $C_1 = 8\sqrt{2} M_\alpha, C_2 = 8\sigma, C_3 = 8\sqrt{2} M_\alpha F_\tau R$.

Proof Note that f_ρ^* can represent a μ-equivalence class in $L^2(\mathcal{X}, \mu)$. When defining the set Ω_0, we actually denote f_ρ^* as the representative $f_\rho^*(x) = \int_{\mathcal{Y}} y \mathrm{d}\rho(y|x)$.

To use Lemma 33, we need to bound the m-th moment of $\xi(x,y) I_{x \in \Omega_0}$.

$$\begin{aligned} \mathbb{E} \left\| \xi(x,y) I_{x \in \Omega_0} \right\|_{\mathcal{H}}^m &= \mathbb{E} \left\| T_\nu^{-\frac{1}{2}} K_x (y - f_\nu(x)) I_{x \in \Omega_0} \right\|_{\mathcal{H}}^m \\ &\leq \mathbb{E} \Big(\left\| T_\nu^{-\frac{1}{2}} k(x, \cdot) \right\|_{\mathcal{H}}^m \mathbb{E} \big(|(y - f_\nu(x)) I_{x \in \Omega_0}|^m \mid x \big) \Big). \end{aligned} \tag{26}$$

Using the inequality $(a+b)^m \leq 2^{m-1}(a^m + b^m)$, we have

$$\begin{aligned} |y - f_\nu(x)|^m &\leq 2^{m-1} \left(\left| f_\nu(x) - f_\rho^*(x) \right|^m + \left| f_\rho^*(x) - y \right|^m \right) \\ &= 2^{m-1} \left(\left| f_\nu(x) - f_\rho^*(x) \right|^m + |\epsilon|^m \right). \end{aligned} \tag{27}$$

Plugging (27) into (26), we have

$$\mathbb{E} \left\| \xi(x,y) I_{x \in \Omega_0} \right\|_{\mathcal{H}}^m \leq 2^{m-1} \mathbb{E} \Big(\left\| T_\nu^{-\frac{1}{2}} k(x, \cdot) \right\|_{\mathcal{H}}^m \left| (f_\nu(x) - f_\rho^*(x)) I_{x \in \Omega_0} \right|^m \Big) \tag{28}$$

$$+ 2^{m-1} \mathbb{E} \Big(\left\| T_\nu^{-\frac{1}{2}} k(x, \cdot) \right\|_{\mathcal{H}}^m \mathbb{E} \big(|\epsilon \ I_{x \in \Omega_0}|^m \mid x \big) \Big) \tag{29}$$

Now we begin to bound (29). Note that we have proved in Lemma 8 that for μ-almost $x \in \mathcal{X}$,

$$\left\| T_\nu^{-\frac{1}{2}} k(x, \cdot) \right\|_{\mathcal{H}} \leq M_\alpha \nu^{\frac{\alpha}{2}};$$

In addition, we also have

$$\begin{aligned} \mathbb{E} \left\| T_\nu^{-\frac{1}{2}} k(x, \cdot) \right\|_{\mathcal{H}}^2 &= \mathbb{E} \Big\| \sum_{i \in N} \big(\frac{1}{\lambda_i + \nu^{-1}} \big)^{\frac{1}{2}} \lambda_i e_i(x) e_i(\cdot) \Big\|_{\mathcal{H}}^2 \\ &= \mathbb{E} \Big(\sum_{i \in N} \frac{\lambda_i}{\lambda_i + \nu^{-1}} e_i^2(x) \Big) \\ &= \sum_{i \in N} \frac{\lambda_i}{\lambda_i + \nu^{-1}} \\ &= \mathcal{N}(\nu). \end{aligned}$$

So we have

$$\mathbb{E}\left\|T_\nu^{-\frac{1}{2}}k(x,\cdot)\right\|_{\mathcal{H}}^m \le \left(M_\alpha\nu^{\frac{\alpha}{2}}\right)^{m-2}\cdot\mathbb{E}\left\|T_\nu^{-\frac{1}{2}}k(x,\cdot)\right\|_{\mathcal{H}}^2 = \left(M_\alpha\nu^{\frac{\alpha}{2}}\right)^{m-2}\mathcal{N}(\nu).$$

Using Assumption 4, we have

$$\mathbb{E}\left(|\epsilon I_{x\in\Omega_0}|^m \mid x\right) \le \mathbb{E}\left(|\epsilon|^m \mid x\right) \le \frac{1}{2}m!\sigma^2L^{m-2}, \quad \mu\text{-a.e. } x\in\mathcal{X},$$

so we get the upper bound of (29), i.e.,

$$(29) \le \frac{1}{2}m!\left(\sqrt{2}\sigma\mathcal{N}^{\frac{1}{2}}(\nu)\right)^2(2M_\alpha\nu^{\frac{\alpha}{2}}L)^{m-2}.$$

Now we begin to bound (28).

(1) When $s\le\alpha_0$, using the definition of Ω_0 and Lemma 7, we have

$$\|(f_\nu - f_\rho^*)I_{x\in\Omega_0}\|_{L^\infty} \le \|f_\nu\|_{L^\infty} + \|f_\rho^*I_{x\in\Omega_0}\|_{L^\infty} \le M_\alpha ER\nu^{\frac{\alpha-s}{2}} + t. \tag{30}$$

(2) When $s>\alpha_0$, without loss of generality, we assume $\alpha_0<\alpha\le s$. using Theorem 6 for $\gamma=\alpha$, we have

$$\|(f_\nu - f_\rho^*)I_{x\in\Omega_0}\|_{L^\infty} \le M_\alpha\|f_\nu - f_\rho^*\|_{[\mathcal{H}]^\alpha} \le M_\alpha F_\tau R\nu^{\frac{\alpha-s}{2}}. \tag{31}$$

Therefore, (30) and (31) imply that for all $0<s\le 2$ we have

$$\|(f_\nu - f_\rho^*)I_{x\in\Omega_0}\|_{L^\infty} \le M_\alpha(E+F_\tau)R\nu^{\frac{\alpha-s}{2}} + t := M. \tag{32}$$

In addition, using Theorem 6 for $\gamma=0$, we also have

$$\mathbb{E}|(f_\nu(x) - f_\rho^*(x))I_{x\in\Omega_0}|^2 \le \mathbb{E}|f_\nu(x) - f_\rho^*(x)|^2 \le (F_\tau R\nu^{-\frac{s}{2}})^2.$$

So we get the upper bound of (28), i.e.,

$$\begin{aligned}(28) &\le 2^{m-1}(M_\alpha\nu^{\frac{\alpha}{2}})^m\cdot\|(f_\nu - f_\rho^*)I_{x\in\Omega_0}\|_{L^\infty}^{m-2}\cdot\mathbb{E}|(f_\nu(x) - f_\rho^*(x))I_{x\in\Omega_0}|^2\\ &\le 2^{m-1}(M_\alpha\nu^{\frac{\alpha}{2}})^m\cdot M^{m-2}\cdot(F_\tau R\nu^{-\frac{s}{2}})^2\\ &\le \frac{1}{2}m!\big(2M_\alpha\nu^{\frac{\alpha}{2}}M\big)^{m-2}\big(2M_\alpha F_\tau R\nu^{\frac{\alpha-s}{2}}\big)^2.\end{aligned}$$

Denote

$$\begin{aligned}\tilde{L} &= 2M_\alpha(M+L)\nu^{\frac{\alpha}{2}}\\ \tilde{\sigma} &= 2M_\alpha F_\tau R\nu^{\frac{\alpha-s}{2}} + \sqrt{2}\sigma\mathcal{N}^{\frac{1}{2}}(\nu),\end{aligned}$$

then the bounds of (28) and (29) show that $\mathbb{E}\,\|\xi(x,y)I_{x\in\Omega_0}\|_{\mathcal{H}}^m \le \frac{1}{2}m!\tilde{\sigma}^2\tilde{L}^{m-2}$. Using Lemma 33, we finish the proof. ∎

Remark 14 *In fact, when we later applying Theorem 13 in the proof of Theorem 15, the truncation in this theorem is necessary only in the $s \leq \alpha_0$ case. But for the completeness and consistency of our proof, we also include $s > \alpha_0$ in this theorem.*

Based on Theorem 13, the following theorem will give an upper bound of

$$\left\| T_\nu^{-\frac{1}{2}} \left[(g_Z - T_X f_\nu) - (g - T f_\nu) \right] \right\|_{\mathcal{H}}.$$

Theorem 15 *Suppose that Assumption 1, 2, 3 and 4 hold for $0 < s \leq 2\tau$ and $\frac{1}{\beta} \leq \alpha_0 < 1$.*

- *In the case of $s + \frac{1}{\beta} > \alpha_0$, by choosing $\nu \asymp n^{\frac{\beta}{s\beta+1}}$, for any fixed $\delta \in (0,1)$, when n is sufficiently large, with probability at least $1 - \delta$, we have*

$$\left\| T_\nu^{-\frac{1}{2}} \left[(g_Z - T_X f_\nu) - (g - T f_\nu) \right] \right\|_{\mathcal{H}} \leq \ln \frac{2}{\delta} C \frac{\nu^{\frac{1}{2\beta}}}{\sqrt{n}} = \ln \frac{2}{\delta} C n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}}, \tag{33}$$

  where C is a constant independent of n and δ.

- *In the case of $s + \frac{1}{\beta} \leq \alpha_0$, for any $\alpha_0 < \alpha \leq 1$, by choosing $\nu \asymp (\frac{n}{\ln^r(n)})^{\frac{1}{\alpha}}$, for some $r > 1$, for any fixed $\delta \in (0,1)$, when n is sufficiently large, with probability at least $1 - \delta$, we have*

$$\left\| T_\nu^{-\frac{1}{2}} \left[(g_Z - T_X f_\nu) - (g - T f_\nu) \right] \right\|_{\mathcal{H}} \leq \ln \frac{2}{\delta} C \frac{\nu^{\frac{\alpha - s}{2}}}{\sqrt{n}} \leq \ln \frac{2}{\delta} C \left(\frac{n}{\ln^r(n)} \right)^{-\frac{1}{2}\frac{s}{\alpha}}, \tag{34}$$

  where C is a constant independent of n and δ.

Proof

The $s + \frac{1}{\beta} > \alpha_0$ case: Denote $\xi_i = \xi(x_i, y_i) = T_\nu^{-\frac{1}{2}}(K_{x_i} y_i - T_{x_i} f_\nu)$, (33) is equivalent to

$$\left\| \frac{1}{n} \sum_{i=1}^n \xi_i - \mathbb{E}\xi_x \right\|_{\mathcal{H}} \leq \ln \frac{2}{\delta} C \frac{\nu^{\frac{1}{2\beta}}}{\sqrt{n}} = \ln \frac{2}{\delta} C n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}}. \tag{35}$$

Consider the subset $\Omega_1 = \{x \in \Omega : |f_\rho^*(x)| \leq t\}$ and $\Omega_2 = \mathcal{X} \backslash \Omega_1$, where t will be chosen appropriately later. Assume that for some $q \geq 2$,

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

Then Assumption 3 shows that there exists $0 < C_q < \infty$ such that $\|f_\rho^*\|_{L^q(\mathcal{X},\mu)} \leq C_q$. Using the Markov inequality, we have

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

Decompose ξ_i as $\xi_i I_{x_i \in \Omega_1} + \xi_i I_{x_i \in \Omega_2}$ and we have

$$\left\| \frac{1}{n} \sum_{i=1}^n \xi_i - \mathbb{E}\xi_x \right\|_{\mathcal{H}} \leq \left\| \frac{1}{n} \sum_{i=1}^n \xi_i I_{x_i \in \Omega_1} - \mathbb{E}\xi_x I_{x \in \Omega_1} \right\|_{\mathcal{H}} + \| \frac{1}{n} \sum_{i=1}^n \xi_i I_{x_i \in \Omega_2} \|_{\mathcal{H}} + \| \mathbb{E}\xi_x I_{x \in \Omega_2} \|_{\mathcal{H}}. \tag{36}$$

For the first term in (36), denoted as I, Theorem 13 shows that there exists $\alpha_0 < \alpha' < s + \frac{1}{\beta}$ such that with probability at least $1-\delta$, we have

$$\mathrm{I} \le \ln \frac{2}{\delta} \left(\frac{C_1 \nu^{\frac{\alpha'}{2}}}{n} \cdot \tilde{M} + \frac{C_2 \nu^{\frac{1}{2\beta}}}{\sqrt{n}} + \frac{C_3 \nu^{\frac{\alpha'-s}{2}}}{\sqrt{n}} \right), \tag{37}$$

where $\tilde{M} = M_{\alpha'}(E + F_\tau) R \nu^{\frac{\alpha'-s}{2}} + t + L$. Recalling that $\mathcal{N}(\nu) \asymp \nu^{\frac{1}{\beta}}$, simple calculation shows that by choosing $\nu = n^{\frac{\beta}{s\beta+1}}$,

- the second term in (37):

$$\frac{C_2 \mathcal{N}^{\frac{1}{2}}(\nu)}{\sqrt{n}} \asymp \frac{\nu^{\frac{1}{2\beta}}}{\sqrt{n}} = n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}}; \tag{38}$$

- the third term in (37):

$$\frac{C_3 \nu^{\frac{\alpha'-s}{2}}}{\sqrt{n}} \asymp n^{\frac{1}{2}(\frac{\alpha'}{s+1/\beta}-1)} \cdot n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}} \lesssim n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}}; \tag{39}$$

- the first term in (37):

$$\frac{C_1 \nu^{\frac{\alpha'}{2}}}{n} \cdot \tilde{M} \asymp \frac{\nu^{\frac{\alpha'}{2}}}{n} \nu^{\frac{\alpha'-s}{2}} + \frac{\nu^{\frac{\alpha'}{2}}}{n} \cdot t + \frac{\nu^{\frac{\alpha'}{2}}}{n} \cdot L. \tag{40}$$

Further calculations show that

$$\frac{\nu^{\frac{\alpha'}{2}}}{n} \nu^{\frac{\alpha'-s}{2}} = n^{\frac{\alpha'}{s+1/\beta}-1} \cdot n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}} \lesssim n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}},$$

and

$$\frac{\nu^{\frac{\alpha'}{2}}}{n} = n^{\frac{1}{2}\frac{\alpha'\beta-s\beta-2}{s\beta+1}} \cdot n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}} \lesssim n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}}.$$

To make (40) $\lesssim n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}}$ when $\nu = n^{\frac{\beta}{s\beta+1}}$, letting $\frac{\nu^{\frac{\alpha'}{2}}}{n} \cdot t \le n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}}$, we have the first restriction of t:

$$\textbf{(R1)}: \quad t \le n^{\frac{1}{2}(1+\frac{1-\alpha'\beta}{s\beta+1})}. \tag{41}$$

That is to say, if we choose $t \le n^{\frac{1}{2}(1+\frac{1-\alpha'\beta}{s\beta+1})}$, we have

$$\mathrm{I} \le \ln \frac{2}{\delta} C \frac{\nu^{\frac{1}{2\beta}}}{\sqrt{n}} = \ln \frac{2}{\delta} C n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}}.$$

For the second term in (36), denoted as II, we have

$$\begin{aligned}
\tau_n := P(\mathrm{II} > \frac{\nu^{\frac{1}{2\beta}}}{\sqrt{n}}) &\le P\Big(\exists x_i \text{ s.t. } x_i \in \Omega_2, \Big) = 1 - P\Big(x_i \notin \Omega_2, \forall x_i, i = 1, 2, \cdots, n \Big) \\
&= 1 - P\Big(x \notin \Omega_2 \Big)^n \\
&= 1 - P\Big(|f_\rho^*(x)| \le t \Big)^n \\
&\le 1 - \Big(1 - \frac{(C_q)^q}{t^q} \Big)^n.
\end{aligned}$$

Letting $\tau_n := P(\mathrm{II} > \frac{\nu^{\frac{1}{2\beta}}}{\sqrt{n}}) \to 0$, we have $t^q \gg n$, i.e. $t \gg n^{\frac{1}{q}}$. This gives the second restriction of t, i.e.,

$$\textbf{(R2)}: \quad t \gg n^{\frac{1}{q}}, \text{ or } n^{\frac{1}{q}} = o(t). \tag{42}$$

For the third term in (36), denoted as III. Since Lemma 8 implies that $\|T_\nu^{-\frac{1}{2}} k(x,\cdot)\|_{\mathcal{H}} \leq M_{\alpha'} \nu^{\frac{\alpha'}{2}}$, μ-a.e. $x \in \mathcal{X}$, so

$$\begin{aligned}
\mathrm{III} &\leq \mathbb{E}\|\xi_x I_{x\in\Omega_2}\|_{\mathcal{H}} \leq \mathbb{E}\Big[\|T_\nu^{-\frac{1}{2}} k(x,\cdot)\|_{\mathcal{H}} \cdot \big|\big(y - f_\nu(x)\big) I_{x\in\Omega_2}\big|\Big] \\
&\leq M_{\alpha'} \nu^{\frac{\alpha'}{2}} \mathbb{E}\big|\big(y - f_\nu(x)\big) I_{x\in\Omega_2}\big| \\
&\leq M_{\alpha'} \nu^{\frac{\alpha'}{2}} \Big(\mathbb{E}\big|\big(f_\rho^*(x) - f_\nu(x)\big) I_{x\in\Omega_2}\big| + \mathbb{E}\big|\big(f_\rho^*(x) - y\big) I_{x\in\Omega_2}\big|\Big) \\
&\leq M_{\alpha'} \nu^{\frac{\alpha'}{2}} \Big(\mathbb{E}\big|\big(f_\rho^*(x) - f_\nu(x)\big) I_{x\in\Omega_2}\big| + \mathbb{E}\big|\epsilon \cdot I_{x\in\Omega_2}\big|\Big).
\end{aligned} \tag{43}$$

Using Cauchy-Schwarz and the bound of approximation error (Theorem 6), we have

$$\mathbb{E}\big|\big(f_\rho^*(x) - f_\nu(x)\big) I_{x\in\Omega_2}\big| \leq \Big(\big\|f_\rho^* - f_\nu\big\|_{L^2}\Big)^{\frac{1}{2}} \cdot \big(P(x \in \Omega_2)\big)^{\frac{1}{2}} \leq R\nu^{-\frac{s}{2}} C_q^{\frac{q}{2}} t^{-\frac{q}{2}}. \tag{44}$$

In addition, we have

$$\mathbb{E}\big|\epsilon \cdot I_{x\in\Omega_2}\big| = \mathbb{E}\left(\mathbb{E}\big|\epsilon \cdot I_{x\in\Omega_2}\big| \;\Big|\; x\right) \leq \sigma\mathbb{E}\,|I_{x\in\Omega_2}| \leq \sigma (C_q)^q t^{-q}. \tag{45}$$

Plugging (44) and (45) into (43), we have

$$\mathrm{III} \leq M_{\alpha'} R C_q^{\frac{q}{2}} \nu^{\frac{\alpha'-s}{2}} t^{-\frac{q}{2}} + M_{\alpha'} \sigma (C_q)^q \nu^{\frac{\alpha'}{2}} t^{-q}. \tag{46}$$

Comparing (46) with $C_3 \frac{\nu^{\frac{\alpha'-s}{2}}}{\sqrt{n}}$ and $C_1 \frac{\nu^{\frac{\alpha'}{2}}}{n}$ in (37). We know that if $t \geq n^{\frac{1}{q}}$, (43) $\leq C\frac{\nu^{\frac{1}{2\beta}}}{\sqrt{n}} = Cn^{-\frac{1}{2}\frac{s\beta}{s\beta+1}}$. So the third term will not give further restriction of t.

To sum up, if we choose t such that restrictions (41) and (42) are satisfied, then we can prove that (35) is satisfied with probability at least $1 - \delta - \tau_n, (\tau_n \to 0)$. Since for a fixed $\delta \in (0,1)$, when n is sufficiently large, τ_n is sufficiently small such that, e.g., $\tau_n < \frac{\delta}{10}$. Without loss of generality, we say (35) is satisfied with probability at least $1 - \delta$.

Recalling restrictions (41) and (42), such t exists if and only if $[\mathcal{H}]^s \hookrightarrow L^q(\mathcal{X}, \mu)$ for some q satisfying

$$\frac{1}{q} < \frac{1}{2}\Big(1 + \frac{1-\alpha'\beta}{s\beta+1}\Big) \Longleftrightarrow q > \frac{2(s\beta+1)}{2 + (s-\alpha')\beta}. \tag{47}$$

If $s > \alpha_0$, $[\mathcal{H}]^s \hookrightarrow L^\infty(\mathcal{X}, \mu)$, hence (47) holds naturally. If $s \leq \alpha_0$, Theorem 5 shows that there exists $\alpha_0 < \alpha'' < \alpha' < s + \frac{1}{\beta}$ such that

$$[\mathcal{H}]^s \hookrightarrow L^{q_s}(\mathcal{X}, \mu), \quad q_s = \frac{2\alpha''}{\alpha'' - s}.$$

Further, $\alpha' > \alpha''$ and $s + \frac{1}{\beta} > \alpha'$ imply that

$$\frac{2\alpha''}{\alpha'' - s} > \frac{2\alpha'}{\alpha' - s} > \frac{2(s\beta+1)}{2 + (s-\alpha')\beta}.$$

So (47) holds for all $s+\frac{1}{\beta}>\alpha_0$ and we finish the proof of this case.

The $s+\frac{1}{\beta}\leq \alpha_0$ case: Denote $\xi_i=\xi(x_i,y_i)=T_\nu^{-\frac{1}{2}}(K_{x_i}y_i-T_{x_i}f_\nu)$, for any fixed $\alpha_0<\alpha\leq 1$, (34) is equivalent to

$$\left\|\frac{1}{n}\sum_{i=1}^n\xi_i-\mathbb{E}\xi_x\right\|_{\mathcal{H}}\leq \ln\frac{2}{\delta}C\frac{\nu^{\frac{\alpha-s}{2}}}{\sqrt{n}}\leq \ln\frac{2}{\delta}C\left(\frac{n}{\ln^r(n)}\right)^{-\frac{1}{2}\frac{s}{\alpha}}, \tag{48}$$

We also consider the subset $\Omega_1=\{x\in\Omega:|f_\rho^*(x)|\leq t\}$ and $\Omega_2=\mathcal{X}\backslash\Omega_1$. Assume that for some $q\geq 2$,

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

Similarly, decompose ξ_i as $\xi_i I_{x_i\in\Omega_1}+\xi_i I_{x_i\in\Omega_2}$ and we have

$$\left\|\frac{1}{n}\sum_{i=1}^n\xi_i-\mathbb{E}\xi_x\right\|_{\mathcal{H}}\leq \left\|\frac{1}{n}\sum_{i=1}^n\xi_i I_{x_i\in\Omega_1}-\mathbb{E}\xi_x I_{x\in\Omega_1}\right\|_{\mathcal{H}}+\|\frac{1}{n}\sum_{i=1}^n\xi_i I_{x_i\in\Omega_2}\|_{\mathcal{H}}+\|\mathbb{E}\xi_x I_{x\in\Omega_2}\|_{\mathcal{H}}. \tag{49}$$

For the first term in (49), denoted as I, Theorem 13 shows that there for this $\alpha>\alpha_0$, with probability at least $1-\delta$, we have

$$\mathrm{I}\leq \ln\frac{2}{\delta}\left(\frac{C_1\nu^{\frac{\alpha}{2}}}{n}\cdot\tilde{M}+\frac{C_2\nu^{\frac{1}{2\beta}}}{\sqrt{n}}+\frac{C_3\nu^{\frac{\alpha-s}{2}}}{\sqrt{n}}\right), \tag{50}$$

where $\tilde{M}=M_\alpha(E+F_\tau)R\nu^{\frac{\alpha-s}{2}}+t+L$. Simple calculation shows that by choosing $\nu=(\frac{n}{\ln^r(n)})^{\frac{1}{\alpha}}$,

- the second term in (50):

$$\frac{C_2\mathcal{N}^{\frac{1}{2}}(\nu)}{\sqrt{n}}\asymp\frac{\nu^{\frac{1}{2\beta}}}{\sqrt{n}}\lesssim\frac{\nu^{\frac{\alpha-s}{2}}}{\sqrt{n}}; \tag{51}$$

- the third term in (50):

$$\frac{C_3\nu^{\frac{\alpha-s}{2}}}{\sqrt{n}}\asymp n^{-\frac{1}{2}}n^{\frac{1}{2}-\frac{s}{2\alpha}}\left(\frac{1}{\ln^r(n)}\right)^{\frac{1}{2}-\frac{s}{2\alpha}}\lesssim\left(\frac{n}{\ln^r(n)}\right)^{-\frac{1}{2}\frac{s}{\alpha}}. \tag{52}$$

- the first term in (50):

$$\frac{C_1\nu^{\frac{\alpha}{2}}}{n}\cdot\tilde{M}\asymp\frac{\nu^{\frac{\alpha}{2}}}{n}\nu^{\frac{\alpha-s}{2}}+\frac{\nu^{\frac{\alpha}{2}}}{n}\cdot t+\frac{\nu^{\frac{\alpha}{2}}}{n}\cdot L. \tag{53}$$

Further calculations show that

$$\frac{\nu^{\frac{\alpha}{2}}}{n}\nu^{\frac{\alpha-s}{2}}=\frac{\nu^{\frac{\alpha-s}{2}}}{\sqrt{n}}\cdot\frac{\nu^{\frac{\alpha}{2}}}{\sqrt{n}}=\frac{\nu^{\frac{\alpha-s}{2}}}{\sqrt{n}}\cdot\left(\frac{1}{\ln^r(n)}\right)^{\frac{\alpha}{2}}\lesssim\frac{\nu^{\frac{\alpha-s}{2}}}{\sqrt{n}}.$$

and

$$\frac{\nu^{\frac{\alpha}{2}}}{n} \lesssim \frac{\nu^{\frac{\alpha-s}{2}}}{\sqrt{n}},$$

To make (53) $\lesssim \left(\frac{n}{\ln^r(n)}\right)^{-\frac{1}{2}\frac{s}{\alpha}}$ when $\nu = (\frac{n}{\ln^r(n)})^{\frac{1}{\alpha}}$, letting $\frac{\nu^{\frac{\alpha}{2}}}{n} \cdot t \leq \frac{\nu^{\frac{\alpha-s}{2}}}{\sqrt{n}}$, we have the first restriction of t (ignoring the log term):

$$\textbf{(R1-2)}: \quad t \leq n^{\frac{1}{2}\left(1-\frac{s}{\alpha}\right)}. \tag{54}$$

For the second and third terms in (49), we repeat the procedure as the case $s + \frac{1}{\beta} > \alpha_0$, therefore the other restriction of t remains unchanged, i.e.,

$$\textbf{(R2)}: \quad t \gg n^{\frac{1}{q}}, \text{ or } n^{\frac{1}{q}} = o(t). \tag{55}$$

These restrictions (54) and (55) shows that such t exists if and only if $[\mathcal{H}]^s \hookrightarrow L^q(\mathcal{X}, \mu)$ for some q satisfying

$$\frac{1}{q} < \frac{1}{2}\left(1 - \frac{s}{\alpha}\right) \Longleftrightarrow q > \frac{2\alpha}{\alpha - s}. \tag{56}$$

Recalling that $\alpha > \alpha_0$ and $s + \frac{1}{\beta} \leq \alpha_0$ implies $s \leq \alpha_0$, Theorem 5 shows that there exists $\alpha_0 < \alpha' < \alpha$ such that

$$[\mathcal{H}]^s \hookrightarrow L^{q_s}(\mathcal{X}, \mu), \quad q_s = \frac{2\alpha'}{\alpha' - s},$$

and

$$\frac{2\alpha'}{\alpha' - s} > \frac{2\alpha}{\alpha - s}.$$

So (56) holds for all $s + \frac{1}{\beta} \leq \alpha_0$ and we finish the proof of this case. ■

Theorem 16 (bound of estimation error) *Suppose that Assumption 1,2, 3 and 4 hold for $0 < s \leq 2\tau$ and $\frac{1}{\beta} \leq \alpha_0 < 1$. Let $\hat{f}_\nu$ be the estimator defined by (11). Then for $0 \leq \gamma \leq 1$ with $\gamma \leq s$:*

- *In the case of $s + \frac{1}{\beta} > \alpha_0$, by choosing $\nu \asymp n^{\frac{\beta}{s\beta+1}}$, for any fixed $\delta \in (0,1)$, when n is sufficiently large, with probability at least $1-\delta$, we have*

$$\left\|\hat{f}_\nu - f_\nu\right\|^2_{[\mathcal{H}]^\gamma} \leq \left(\ln\frac{6}{\delta}\right)^2 C n^{-\frac{(s-\gamma)\beta}{s\beta+1}}, \tag{57}$$

  where C is a constant independent of n and δ.

- *In the case of $s + \frac{1}{\beta} \leq \alpha_0$, for any $\alpha_0 < \alpha \leq 1$, by choosing $\nu \asymp (\frac{n}{\ln^r(n)})^{\frac{1}{\alpha}}$, for some $r > 1$, for any fixed $\delta \in (0,1)$, when n is sufficiently large, with probability at least $1-\delta$, we have*

$$\left\|\hat{f}_\nu - f_\nu\right\|^2_{[\mathcal{H}]^\gamma} \leq \left(\ln\frac{6}{\delta}\right)^2 C \left(\frac{n}{\ln^r(n)}\right)^{-\frac{s-\gamma}{\alpha}}, \tag{58}$$

  where C is a constant independent of n and δ.

Proof Using Lemma 12, Theorem 15 and Lemma 10 for $\frac{\delta}{3} \in (0, \frac{1}{3})$, with probability at least $1-\delta$, we have the following results hold simultaneously

$$\left\| T_\nu^{-\frac{1}{2}} T_{X\nu}^{\frac{1}{2}} \right\|^2 \leq 2, \quad \left\| T_\nu^{\frac{1}{2}} T_{X\nu}^{-\frac{1}{2}} \right\|^2 \leq 3; \tag{59}$$

(33) and (34);

$$\|T_X - T\| \leq \frac{8\sqrt{2}\kappa^2}{\sqrt{n}} \ln \frac{6}{\delta}. \tag{60}$$

Note that when choosing ν as in (57) or (58), the condition (25) required in Lemma 12 is always satisfied when n is sufficiently large.

Step 1: First, we rewrite the estimation error as follows,

$$\begin{aligned}
\left\| \hat{f}_\nu - f_\nu \right\|_{[\mathcal{H}]^\gamma} &= \left\| L_k^{-\frac{\gamma}{2}} S_k \left(\hat{f}_\nu - f_\nu \right) \right\|_{L^2} \\
&= \left\| L_k^{-\frac{\gamma}{2}} S_k T_\nu^{-\frac{1}{2}} \cdot T_\nu^{\frac{1}{2}} T_{X\nu}^{-\frac{1}{2}} \cdot T_{X\nu}^{\frac{1}{2}} \left(\hat{f}_\nu - f_\nu \right) \right\|_{L^2} \\
&\leq \left\| L_k^{-\frac{\gamma}{2}} S_k T_\nu^{-\frac{1}{2}} \right\|_{\mathscr{B}(\mathcal{H}, L^2)} \cdot \left\| T_\nu^{\frac{1}{2}} T_{X\nu}^{-\frac{1}{2}} \right\|_{\mathscr{B}(\mathcal{H})} \cdot \left\| T_{X\nu}^{\frac{1}{2}} \left(\hat{f}_\nu - f_\nu \right) \right\|_{\mathcal{H}}.
\end{aligned} \tag{61}$$

For any $f \in \mathcal{H}$ and $\|f\|_{\mathcal{H}} = 1$, suppose that $f = \sum_{i \in N} a_i \lambda_i^{1/2} e_i$ satisfying that $\sum_{i \in N} a_i^2 = 1$. So for the first term in (61), we have

$$\begin{aligned}
\left\| L_k^{-\frac{\gamma}{2}} S_k T_\nu^{-\frac{1}{2}} \right\|_{\mathscr{B}(\mathcal{H}, L^2)} &= \sup_{\|f\|_{\mathcal{H}}=1} \left\| L_k^{-\frac{\gamma}{2}} S_k T_\nu^{-\frac{1}{2}} f \right\|_{L^2} \\
&\leq \sup_{\|f\|_{\mathcal{H}}=1} \left\| \sum_{i \in N} \frac{\lambda_i^{\frac{1-\gamma}{2}}}{(\lambda_i + \nu^{-1})^{\frac{1}{2}}} a_i e_i \right\|_{L^2} \\
&\leq \sup_{i \in N} \frac{\lambda_i^{\frac{1-\gamma}{2}}}{(\lambda_i + \nu^{-1})^{\frac{1}{2}}} \cdot \sup_{\|f\|_{\mathcal{H}}=1} \left\| \sum_{i \in N} a_i e_i \right\|_{L^2} \\
&\leq \nu^{\frac{\gamma}{2}},
\end{aligned}$$

where we use Lemma 30 for the last inequality. For the second term in (61), (59) shows that

$$\left\| T_\nu^{\frac{1}{2}} T_{X\nu}^{-\frac{1}{2}} \right\|_{\mathscr{B}(\mathcal{H})} \leq 3.$$

For the third term in (61), noticing that $z\varphi_\nu + \psi_\nu = 1$, we have

$$\begin{aligned}
\hat{f}_\nu - f_\nu &= \varphi_\nu(T_X) g_Z - (T_X \varphi_\nu(T_X) + \psi_\nu(T_X)) f_\nu \\
&= \varphi_\nu(T_X)(g_Z - T_X f_\nu) - \psi_\nu(T_X) f_\nu
\end{aligned}$$

So for the third term in (61),

$$\left\|T_{X\nu}^{\frac{1}{2}}\left(\hat{f}_\nu - f_\nu\right)\right\|_{\mathcal{H}} \leq \left\|T_{X\nu}^{\frac{1}{2}}\varphi_\nu\left(T_X\right)\left(g_Z - T_X f_\nu\right)\right\|_{\mathcal{H}} + \left\|T_{X\nu}^{\frac{1}{2}}\psi_\nu\left(T_X\right) f_\nu\right\|_{\mathcal{H}}. \tag{62}$$

Step 2: Now we begin to bound the first term in (62), i.e.,

$$\begin{aligned}\left\|T_{X\nu}^{\frac{1}{2}}\varphi_\nu\left(T_X\right)\left(g_Z - T_X f_\nu\right)\right\|_{\mathcal{H}} &= \left\|T_{X\nu}^{\frac{1}{2}}\varphi_\nu\left(T_X\right)T_{X\nu}^{\frac{1}{2}}\cdot T_{X\nu}^{-\frac{1}{2}}T_\nu^{\frac{1}{2}}\cdot T_\nu^{-\frac{1}{2}}\left(g_Z - T_X f_\nu\right)\right\|_{\mathcal{H}} \\ &\leq \left\|T_{X\nu}^{\frac{1}{2}}\varphi_\nu\left(T_X\right)T_{X\nu}^{\frac{1}{2}}\right\|_{\mathscr{B}(\mathcal{H})}\cdot\left\|T_{X\nu}^{-\frac{1}{2}}T_\nu^{\frac{1}{2}}\right\|_{\mathscr{B}(\mathcal{H})}\cdot\left\|T_\nu^{-\frac{1}{2}}\left(g_Z - T_X f_\nu\right)\right\|_{\mathcal{H}}.\end{aligned} \tag{63}$$

The property of filter function (9) shows that $z\varphi_\nu(z) \leq E$ and $\varphi_\nu(z) \leq E\nu$. So we have

$$\left\|T_{X\nu}^{\frac{1}{2}}\varphi_\nu\left(T_X\right)T_{X\nu}^{\frac{1}{2}}\right\|_{\mathscr{B}(\mathcal{H})} = \left\|\left(T_X + \nu^{-1}\right)\varphi_\nu\left(T_X\right)\right\|_{\mathscr{B}(\mathcal{H})} \leq 2E; \tag{64}$$

(59) shows that

$$\left\|T_{X\nu}^{-\frac{1}{2}}T_\nu^{\frac{1}{2}}\right\|_{\mathscr{B}(\mathcal{H})} \leq 2; \tag{65}$$

In addition, recalling that at the beginning we have assumed that (33) and (34) hold, therefore we have

- In the case of $s + \frac{1}{\beta} > \alpha_0$, by choosing $\nu \asymp n^{\frac{\beta}{s\beta+1}}$, we have

$$\begin{aligned}\left\|T_\nu^{-\frac{1}{2}}\left(g_Z - T_X f_\nu\right)\right\|_{\mathcal{H}} &\leq \left\|T_\nu^{-\frac{1}{2}}\left[\left(g_Z - T_X f_\nu\right) - \left(g - T f_\nu\right)\right]\right\|_{\mathcal{H}} + \left\|T_\nu^{-\frac{1}{2}}\left(g - T f_\nu\right)\right\|_{\mathcal{H}} \\ &\leq \ln(\frac{6}{\delta})Cn^{-\frac{1}{2}\frac{s\beta}{s\beta+1}} + \left\|T_\nu^{-\frac{1}{2}}\left(S_k^* f_\rho^* - S_k^* S_k f_\nu\right)\right\|_{\mathcal{H}} \\ &\leq \ln(\frac{6}{\delta})Cn^{-\frac{1}{2}\frac{s\beta}{s\beta+1}} + \left\|T_\nu^{-\frac{1}{2}}S_k^*\right\|_{\mathscr{B}(L^2,\mathcal{H})}\|f_\rho^* - f_\nu\|_{L^2} \\ &\leq \ln(\frac{6}{\delta})Cn^{-\frac{1}{2}\frac{s\beta}{s\beta+1}} + \|f_\rho^* - f_\nu\|_{L^2} \\ &\leq \ln(\frac{6}{\delta})Cn^{-\frac{1}{2}\frac{s\beta}{s\beta+1}} + F_\tau R n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}}. \\ &\leq \ln(\frac{6}{\delta})Cn^{-\frac{1}{2}\frac{s\beta}{s\beta+1}},\end{aligned} \tag{66}$$

  where we use the fact that $\left\|T_\nu^{-\frac{1}{2}}S_k^*\right\|_{\mathscr{B}(L^2,\mathcal{H})} \leq 1$ and use Theorem 6 with $\gamma = 0$ to bound $\|f_\rho^* - f_\nu\|_{L^2}$.

- In the case of $s+\frac{1}{\beta}\leq\alpha_0$, for any $\alpha_0<\alpha\leq 1$, by choosing $\nu\asymp(\frac{n}{\ln^r(n)})^{\frac{1}{\alpha}}$, for some $r>1$, we have

$$\begin{aligned}\left\|T_\nu^{-\frac{1}{2}}\left(g_Z-T_Xf_\nu\right)\right\|_{\mathcal{H}}&\leq\left\|T_\nu^{-\frac{1}{2}}\left[\left(g_Z-T_Xf_\nu\right)-\left(g-Tf_\nu\right)\right]\right\|_{\mathcal{H}}+\left\|T_\nu^{-\frac{1}{2}}\left(g-Tf_\nu\right)\right\|_{\mathcal{H}}\\&\leq\ln\frac{6}{\delta}C\left(\frac{n}{\ln^r(n)}\right)^{-\frac{1}{2}\frac{s}{\alpha}}+F_\tau R\left(\frac{n}{\ln^r(n)}\right)^{-\frac{1}{2}\frac{s}{\alpha}}\\&\leq\ln\frac{6}{\delta}C\left(\frac{n}{\ln^r(n)}\right)^{-\frac{1}{2}\frac{s}{\alpha}}.\end{aligned}\tag{67}$$

Therefore, plugging (64) (65) (66) (67) into (63), we get the desired upper bounds of the first term in (62). Specifically, the bound in (66) determines the bound of (63) in the the case of $s+\frac{1}{\beta}>\alpha_0$; and (67) determines the case of $s+\frac{1}{\beta}\leq\alpha_0$.

Step 3: Now we begin to bound the second term in (62), i.e.,

$$\left\|T_{X\nu}^{\frac{1}{2}}\psi_\nu\left(T_X\right)f_\nu\right\|_{\mathcal{H}}.\tag{68}$$

We discuss three conditions of s.

- $0<s<1$: Since $(a+b)^p\leq a^p+b^p$ for $p\in[0,1]$, we have

$$\left\|T_{X\nu}^{\frac{1}{2}}\psi_\nu\left(T_X\right)\right\|_{\mathscr{B}(\mathcal{H})}\leq\sup_{z\in[0,\kappa^2]}(z+\nu^{-1})^{\frac{1}{2}}\psi_\nu(z)\leq\sup_{z\in[0,\kappa^2]}(z^{\frac{1}{2}}+\nu^{-\frac{1}{2}})\psi_\nu(z).$$

Using the property of filter function (10), we have

$$\sup_{z\in[0,\kappa^2]}(z^{\frac{1}{2}}+\nu^{-\frac{1}{2}})\psi_\nu(z)\leq F_\tau\nu^{-\frac{1}{2}}+\nu^{-\frac{1}{2}}F_\tau\leq 2F_\tau\nu^{-\frac{1}{2}}.\tag{69}$$

Furthermore, using the property of filter function (9) and recalling that

$$f_\nu=\varphi_\nu(T)S_k^*L_k^{\frac{s}{2}}g_0=\varphi_\nu(T)T^{\frac{s}{2}}S_k^*g_0,$$

for some $g_0\in L^2$ with $\|g_0\|_{L^2}\leq R$, we have

$$\begin{aligned}\left\|T_{X\nu}^{\frac{1}{2}}\psi_\nu\left(T_X\right)f_\nu\right\|_{\mathcal{H}}&\leq\left\|T_{X\nu}^{\frac{1}{2}}\psi_\nu\left(T_X\right)\right\|_{\mathscr{B}(\mathcal{H})}\left\|\varphi_\nu(T)T^{\frac{s}{2}}S_k^*g_0\right\|_{\mathcal{H}}\\&\leq 2F_\tau\nu^{-\frac{1}{2}}\cdot\left\|\varphi_\nu(T)T^{\frac{s}{2}}S_k^*\right\|\|g_0\|_{L^2}\\&=2F_\tau\nu^{-\frac{1}{2}}\cdot\left\|\varphi_\nu(T)T^{\frac{s}{2}}T^{\frac{1}{2}}\right\|\|g_0\|_{L^2}\\&\leq 2F_\tau\nu^{-\frac{1}{2}}\cdot\left\|T^{\frac{1+s}{2}}\varphi_\nu(T)\right\|_{\mathscr{B}(\mathcal{H})}\|g_0\|_{L^2}\\&\leq 2F_\tau\nu^{-\frac{1}{2}}E\nu^{\frac{1-s}{2}}R\\&=2F_\tau ER\nu^{-\frac{s}{2}}.\end{aligned}\tag{70}$$

- $1 \leq s \leq 2$: We can rewrite (68) as follows,

$$\begin{aligned}\left\|T_{X\nu}^{\frac{1}{2}}\psi_\nu\left(T_X\right)f_\nu\right\|_{\mathcal{H}} &= \left\|T_{X\nu}^{\frac{1}{2}}\psi_\nu\left(T_X\right)\varphi_\nu(T)T^{\frac{s}{2}}S_k^*g_0\right\|_{\mathcal{H}} \\ &= \left\|T_{X\nu}^{\frac{1}{2}}\psi_\nu\left(T_X\right)\varphi_\nu(T)T^{\frac{s}{2}}S_k^*\right\|\|g_0\|_{L^2} \\ &= \left\|T_{X\nu}^{\frac{1}{2}}\psi_\nu\left(T_X\right)\varphi_\nu(T)T^{\frac{s}{2}}T^{\frac{1}{2}}\right\|\|g_0\|_{L^2} \\ &\leq \left\|T_{X\nu}^{\frac{1}{2}}\psi_\nu\left(T_X\right)\varphi_\nu(T)T^{\frac{s+1}{2}}\right\|R. \end{aligned} \tag{71}$$

Next, we can further decompose (71) as follows

$$\begin{aligned}\left\|T_{X\nu}^{\frac{1}{2}}\psi_\nu\left(T_X\right)\varphi_\nu(T)T^{\frac{s+1}{2}}\right\| &= \left\|T_{X\nu}^{\frac{1}{2}}\psi_\nu\left(T_X\right)T_{X\nu}^{\frac{s-1}{2}}\cdot T_{X\nu}^{-\frac{s-1}{2}}T_\nu^{\frac{s-1}{2}}\cdot T_\nu^{-\frac{s-1}{2}}T^{\frac{s-1}{2}}\cdot T^{-\frac{s-1}{2}}\varphi_\nu(T)T^{\frac{s+1}{2}}\right\| \\ &= \left\|T_{X\nu}^{\frac{1}{2}}\psi_\nu\left(T_X\right)T_{X\nu}^{\frac{s-1}{2}}\right\|\left\|T_{X\nu}^{-\frac{s-1}{2}}T_\nu^{\frac{s-1}{2}}\right\|\left\|T_\nu^{-\frac{s-1}{2}}T^{\frac{s-1}{2}}\right\|\left\|T^{-\frac{s-1}{2}}\varphi_\nu(T)T^{\frac{s+1}{2}}\right\| \\ &= \left\|T_{X\nu}^{\frac{s}{2}}\psi_\nu\left(T_X\right)\right\|\left\|T_{X\nu}^{-\frac{s-1}{2}}T_\nu^{\frac{s-1}{2}}\right\|\left\|T_\nu^{-\frac{s-1}{2}}T^{\frac{s-1}{2}}\right\|\left\|T\varphi_\nu(T)\right\|. \end{aligned} \tag{72}$$

Next, we need to bound the four terms in (72). For the first term in (72), using the inequality $(a+b)^p \leq a^p + b^p$ for $p \in [0,1]$ again, we have

$$\left\|T_{X\nu}^{\frac{s}{2}}\psi_\nu\left(T_X\right)\right\|_{\mathscr{B}(\mathcal{H})} \leq \sup_{z\in[0,\kappa^2]}(z+\nu^{-1})^{\frac{s}{2}}\psi_\nu(z) \leq \sup_{z\in[0,\kappa^2]}(z^{\frac{s}{2}}+\nu^{-\frac{s}{2}})\psi_\nu(z) \leq 2F_\tau\nu^{-\frac{s}{2}}. \tag{73}$$

For the second term in (72), using Lemma 34 and (59), we have,

$$\left\|T_{X\nu}^{-\frac{s-1}{2}}T_\nu^{\frac{s-1}{2}}\right\| \leq \left\|T_{X\nu}^{-\frac{1}{2}}T_\nu^{\frac{1}{2}}\right\|^{s-1} \leq 3^{s-1} \leq 3. \tag{74}$$

For the third term in (72),

$$\left\|T_\nu^{-\frac{s-1}{2}}T^{\frac{s-1}{2}}\right\| = \sup_{i\in N}\left(\frac{\lambda_i}{\lambda_i+\nu^{-1}}\right)^{\frac{s-1}{2}} \leq 1. \tag{75}$$

For the fourth term in (72), using the property of filter function (9), we have

$$\left\|T\varphi_\nu(T)\right\| \leq E. \tag{76}$$

Plugging (73) (74) (75) (76) into (72), we obtain the bound

$$\left\|T_{X\nu}^{\frac{1}{2}}\psi_\nu\left(T_X\right)f_\nu\right\|_{\mathcal{H}} \leq 6EF_\tau R\nu^{-\frac{s}{2}}. \tag{77}$$

- $s > 2$: Recalling (71), we have

$$\begin{aligned}\left\|T_{X\nu}^{\frac{1}{2}}\psi_\nu\left(T_X\right)f_\nu\right\|_{\mathcal{H}} &\le \left\|T_{X\nu}^{\frac{1}{2}}\psi_\nu\left(T_X\right)\varphi_\nu(T)T^{\frac{s+1}{2}}\right\| R \\ &\le \left\|T_{X\nu}^{\frac{1}{2}}\psi_\nu\left(T_X\right)T^{\frac{s-1}{2}}\right\| \|T\varphi_\nu(T)\| R \\ &\le \left\|T_{X\nu}^{\frac{1}{2}}\psi_\nu\left(T_X\right)T^{\frac{s-1}{2}}\right\| ER. \end{aligned} \tag{78}$$

Further, we can have the following decomposition

$$T_{X\nu}^{\frac{1}{2}}\psi_\nu\left(T_X\right)T^{\frac{s-1}{2}} = T_{X\nu}^{\frac{1}{2}}\psi_\nu\left(T_X\right)\left(T^{\frac{s-1}{2}} - T_X^{\frac{s-1}{2}}\right) + T_{X\nu}^{\frac{1}{2}}\psi_\nu\left(T_X\right)T_X^{\frac{s-1}{2}}.$$

So we have

$$\left\|T_{X\nu}^{\frac{1}{2}}\psi_\nu\left(T_X\right)T^{\frac{s-1}{2}}\right\| \le \left\|T_{X\nu}^{\frac{1}{2}}\psi_\nu\left(T_X\right)\right\|\left\|T^{\frac{s-1}{2}} - T_X^{\frac{s-1}{2}}\right\| + \left\|T_{X\nu}^{\frac{1}{2}}\psi_\nu\left(T_X\right)T_X^{\frac{s-1}{2}}\right\|. \tag{79}$$

For the first term in (79), using Lemma 35 and the fact that $\|T_X\|, \|T\| \le \kappa^2$, we have

$$\left\|T^{\frac{s-1}{2}} - T_X^{\frac{s-1}{2}}\right\| \le \begin{cases} \|T - T_X\|^{\frac{s-1}{2}} & s \in (2,3], \\ \frac{s-1}{2}\kappa^{s-3}\|T - T_X\| & s \ge 3. \end{cases} \tag{80}$$

In addition, (60) shows that

$$\|T_X - T\| \le \|T_X - T\|_2 \le \frac{8\sqrt{2}\kappa^2}{\sqrt{n}}\ln\frac{6}{\delta}. \tag{81}$$

Further, recalling (69), we have

$$\left\|T_{X\nu}^{\frac{1}{2}}\psi_\nu\left(T_X\right)\right\| \le 2F_\tau\nu^{-\frac{1}{2}}. \tag{82}$$

In addition, similarly as (73), we have

$$\begin{aligned}\left\|T_{X\nu}^{\frac{1}{2}}\psi_\nu\left(T_X\right)T_X^{\frac{s-1}{2}}\right\| &\le \left\|T_X^{\frac{1}{2}}\psi_\nu\left(T_X\right)T_X^{\frac{s-1}{2}}\right\| + \nu^{-\frac{1}{2}}\left\|\psi_\nu\left(T_X\right)T_X^{\frac{s-1}{2}}\right\| \\ &= \left\|T_X^{\frac{s}{2}}\psi_\nu\left(T_X\right)\right\| + \nu^{-\frac{1}{2}}\left\|T_X^{\frac{s-1}{2}}\psi_\nu\left(T_X\right)\right\| \\ &\le F_\tau\nu^{-\frac{s}{2}} + \nu^{-\frac{1}{2}}F_\tau\nu^{\frac{1-s}{2}} \\ &= 2F_\tau\nu^{-\frac{s}{2}}. \end{aligned} \tag{83}$$

To sum up, denote

$$\Delta_0 := 2EF_\tau R\nu^{-\frac{1}{2}}\kappa^{s-1}\cdot\begin{cases} n^{-\frac{s-1}{4}}\left(8\sqrt{2}\ln\frac{6}{\delta}\right)^{\frac{s-1}{2}} & s \in (2,3], \\ n^{-\frac{1}{2}}\cdot\frac{s-1}{2}\cdot 8\sqrt{2}\ln\frac{6}{\delta}, & s \ge 3. \end{cases}$$

Then plugging (80) $\sim$ (83) into (79) and use (78), we have

$$\left\| T_{X\nu}^{\frac{1}{2}} \psi_\nu \left(T_X \right) f_\nu \right\|_{\mathcal{H}} \leq \Delta_0 + 2EF_\tau R \nu^{-\frac{s}{2}}.$$

Without loss of generality, we assume that $\ln \frac{6}{\delta} > 1$. Simple calculation shows that,

$$\Delta_0 \leq 32 \max\left(\frac{s-1}{2}, 1 \right) EF_\tau R \kappa^{s-1} \nu^{-\frac{1}{2}} n^{-\frac{\min(s,3)-1}{4}} \ln \frac{6}{\delta} := \Delta_1. \tag{84}$$

Then we have

$$\left\| T_{X\nu}^{\frac{1}{2}} \psi_\nu \left(T_X \right) f_\nu \right\|_{\mathcal{H}} \leq \Delta_1 + 2EF_\tau R \nu^{-\frac{s}{2}}. \tag{85}$$

Combining the bounds of three conditions of s, i.e., (70) (77) (85), we finally bound the goal of Step 3, i.e., (68) by

$$\left\| T_{X\nu}^{\frac{1}{2}} \psi_\nu \left(T_X \right) f_\nu \right\|_{\mathcal{H}} \leq 6F_\tau ER \nu^{-\frac{s}{2}} + \Delta_1 I_{s>2}.$$

Step 4: Now we are able to use the results of Step1 $\sim$ Step3 to finish the proof of the estimation error. Still, we consider two cases, $s + \frac{1}{\beta} > \alpha_0$ and $s + \frac{1}{\beta} \leq \alpha_0$.

- $s + \frac{1}{\beta} > \alpha_0$: Plugging the results of Step2 and Step3 into (62) and using the decomposition (61), by choosing $\nu \asymp n^{\frac{\beta}{s\beta+1}}$, we have

$$\begin{aligned} \left\| \hat{f}_\nu - f_\nu \right\|_{[\mathcal{H}]^\gamma} &\leq 3\nu^{\frac{\gamma}{2}} \left(\ln(\frac{6}{\delta}) C n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}} + 6F_\tau ER\nu^{-\frac{s}{2}} + \Delta_1 I_{s>2} \right) \\ &= 3n^{\frac{1}{2}\frac{\gamma\beta}{s\beta+1}} \left(\ln(\frac{6}{\delta}) C n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}} + 6F_\tau ER n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}} + \Delta_1 I_{s>2} \right). \end{aligned}$$

  Recalling the expression of Δ_1 in (84), when $2 < s \leq 3$,

$$\Delta_1 \asymp n^{-\frac{r_0}{2}},$$

  where

$$r_0 = \frac{\beta}{s\beta+1} + \frac{s-1}{2}.$$

  Since $s > 2$ implies $s + \frac{1}{\beta} > 2$, so we have

$$r_0 - \frac{s\beta}{s\beta+1} = \frac{s-1}{2} - \frac{s-1}{s+\frac{1}{\beta}} > 0.$$

  So we have $\Delta_1 \lesssim n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}}$.

When $s > 3$, we also have $r_0 = \frac{\beta}{s\beta+1} + 1 > \frac{s\beta}{s\beta+1}$. Therefore, we know that

$$\Delta_1 I_{s>2} \leq C \ln \frac{6}{\delta} n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}}.$$

To sum up, we prove that when $s + \frac{1}{\beta} > \alpha_0$, the estimation error satisfies that

$$\left\| \hat{f}_\nu - f_\nu \right\|_{[\mathcal{H}]^\gamma} \leq \ln \frac{6}{\delta} C n^{-\frac{1}{2}\frac{(s-\gamma)\beta}{s\beta+1}}. \tag{86}$$

- $s + \frac{1}{\beta} \leq \alpha_0$: In this case, $s \leq 1$. Similarly, for some fixed $\alpha_0 < \alpha \leq 1$, by choosing $\nu \asymp (\frac{n}{\ln^r(n)})^{\frac{1}{\alpha}}$, we have

$$\begin{aligned}
\left\| \hat{f}_\nu - f_\nu \right\|_{[\mathcal{H}]^\gamma} &\leq 3\nu^{\frac{\gamma}{2}} \left(\ln \frac{6}{\delta} C \left(\frac{n}{\ln^r(n)} \right)^{-\frac{1}{2}\frac{s}{\alpha}} + 6F_\tau E R \nu^{-\frac{s}{2}} \right) \\
&= 3 \left(\frac{n}{\ln^r(n)} \right)^{\frac{\gamma}{2\alpha}} \left(\ln \frac{6}{\delta} C \left(\frac{n}{\ln^r(n)} \right)^{-\frac{1}{2}\frac{s}{\alpha}} + 6F_\tau E R \left(\frac{n}{\ln^r(n)} \right)^{-\frac{1}{2}\frac{s}{\alpha}} \right) \\
&\leq \ln \frac{6}{\delta} C \left(\frac{n}{\ln^r(n)} \right)^{-\frac{s-\gamma}{2\alpha}}.
\end{aligned} \tag{87}$$

Then, the proof of Theorem 16 follows from (86) and (87). ■

Proof of Theorem 1 We first decompose the $[\mathcal{H}]^\gamma$-norm generalization error into two terms, which are often referred to as the approximation error and the estimation error:

$$\left\| \hat{f}_\nu - f_\rho^* \right\|_{[\mathcal{H}]^\gamma} = \left\| f_\nu - f_\rho^* \right\|_{[\mathcal{H}]^\gamma} + \left\| \hat{f}_\nu - f_\nu \right\|_{[\mathcal{H}]^\gamma}. \tag{88}$$

For the approximation error, Theorem 6 proves that

- By choosing $\nu \asymp n^{\frac{\beta}{s\beta+1}}$,

$$\left\| f_\nu - f_\rho^* \right\|_{[\mathcal{H}]^\gamma} \leq F_\tau R n^{-\frac{1}{2}\frac{(s-\gamma)\beta}{s\beta+1}}; \tag{89}$$

- by choosing $\nu \asymp (\frac{n}{\ln^r(n)})^{\frac{1}{\alpha}}$, for some $r > 1$,

$$\left\| f_\nu - f_\rho^* \right\|_{[\mathcal{H}]^\gamma} \leq F_\tau R \left(\frac{n}{\ln^r(n)} \right)^{-\frac{s-\gamma}{2\alpha}}. \tag{90}$$

Then the proof follows from plugging (89), (90) and the bounds of estimation error in Theorem 16 into (88).

7.4 Lower bound

The following lemma is a standard approach to derive the minimax lower bound, which can be found in Tsybakov (2009, Theorem 2.5).

Lemma 17 *Suppose that there is a non-parametric class of functions Θ and a (semi-)distance $d(\cdot,\cdot)$ on Θ. $\{P_\theta, \theta \in \Theta\}$ is a family of probability distributions indexed by Θ. Assume that $M \geq 2$ and suppose that Θ contains elements $\theta_0, \theta_1, \cdots, \theta_M$ such that,*

(1) $d(\theta_j, \theta_k) \geq 2s > 0, \quad \forall 0 \leq j < k \leq M$;

(2) $P_j \ll P_0, \quad \forall j = 1, \cdots, M$, *and*

$$\frac{1}{M}\sum_{j=1}^{M} K(P_j, P_0) \leq a \log M,$$

with $0 < a < 1/8$ and $P_j = P_{\theta_j}, j = 0, 1, \cdots, M$. Then

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} P_\theta(d(\hat{\theta}, \theta) \geq s) \geq \frac{\sqrt{M}}{1+\sqrt{M}} \left(1 - 2a - \sqrt{\frac{2a}{\log M}}\right).$$

Lemma 18 *Suppose that μ is a distribution on $\mathcal{X}$ and $f_i \in L^2(\mathcal{X}, \mu)$. Suppose that*

$$y = f_i(x) + \epsilon, \quad i = 1, 2,$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ are independent Gaussian random error. Denote the two corresponding distributions on $\mathcal{X} \times \mathcal{Y}$ as $\rho_i, i = 1, 2$. The KL divergence of two probability distributions on Ω is

$$K(P_1, P_2) \coloneqq \int_\Omega \log\left(\frac{\mathrm{d}P_1}{\mathrm{d}P_2}\right) \mathrm{d}P_1,$$

if $P_1 \ll P_2$ and otherwise $K(P_1, P_2) \coloneqq \infty$. Then we have

$$\mathrm{KL}(\rho_1^n, \rho_2^n) = n\mathrm{KL}(\rho_1, \rho_2) = \frac{n}{2\sigma^2} \|f_1 - f_2\|^2_{L^2(\mathcal{X}, d\mu)},$$

where ρ_i^n denotes the independent product of n distributions $\rho_i, i = 1, 2$.

Proof The lemma directly follows from the definition of KL divergence and the fact that

$$\mathrm{KL}\left(N\left(f_1(x), \sigma^2\right), N\left(f_2(x), \sigma^2\right)\right) = \frac{1}{2\sigma^2} |f_1(x) - f_2(x)|^2.$$

■

The following lemma is a result from Tsybakov (2009, Lemma 2.9)

Lemma 19 *Denote $\Omega = \{\omega = (\omega_1, \cdots, \omega_m), \omega_i \in \{0, 1\}\} = \{0, 1\}^m$. Let $m \geq 8$, there exists a subset $\left\{\omega^{(0)}, \cdots, \omega^{(M)}\right\}$ of Ω such that $\omega^{(0)} = (0, \cdots, 0)$,*

$$d_{Ham}\left(\omega^{(i)}, \omega^{(j)}\right) \coloneqq \sum_{k=1}^{m} \left|\omega_k^{(i)} - \omega_k^{(j)}\right| \geq \frac{m}{8}, \quad \forall 0 \leq i < j \leq M,$$

and $M \geq 2^{m/8}$.

Now we are ready to prove the minimax lower bound given by Theorem 2.

Proof of Theorem 2 We will construct a family of probability distributions on $\mathcal{X}\times\mathcal{Y}$ and apply Lemma 17. Recall that μ is a probability distribution on $\mathcal{X}$ such that Assumption 1 is satisfied. Denote the class of functions

$$B^s(R)=\left\{f\in[\mathcal{H}]^s:\|f\|_{[\mathcal{H}]^s}\le R\right\},$$

and for every $f\in B^s(R)$, define the probability distribution ρ_f on $\mathcal{X}\times\mathcal{Y}$ such that

$$y=f(x)+\epsilon,\quad x\sim\mu,$$

where $\epsilon\sim\mathcal{N}(0,\bar{\sigma}^2)$ and $\bar{\sigma}=\min(\sigma,L)$. It is easy to show that such ρ_f falls into the family $\mathcal{P}$ in Lemma 2. (Assumption 1 and 3 are satisfied obviously. Assumption 4 follows from results of moments of Gaussian random variables, see, e.g., Fischer and Steinwart (2020, Lemma 21)).

Using Lemma 19, for $m=n^{\frac{1}{s\beta+1}}$, there exists $\omega^{(0)},\cdots,\omega^{(M)}\in\{0,1\}^m$ for some $M\ge 2^{m/8}$ such that

$$\sum_{k=1}^m\left|\omega_k^{(i)}-\omega_k^{(j)}\right|\ge\frac{m}{8},\quad\forall 0\le i<j\le M. \tag{91}$$

For $\epsilon=C_0m^{-(s-\gamma)\beta-1}$, define the functions $f_i,i=1,2,\cdots,M$ as

$$f_i:=\epsilon^{1/2}\sum_{k=1}^m\omega_k^{(i)}\lambda_{m+k}^{\frac{\gamma}{2}}e_{m+k}.$$

Since

$$\begin{aligned}\|f_i\|_{[\mathcal{H}]^s}&=\epsilon\sum_{k=1}^m\lambda_{m+k}^{\gamma-s}\left(\omega_k^{(i)}\right)^2\le\epsilon\sum_{k=1}^m\lambda_{2m}^{\gamma-s}\\&\le 2^{(s-\gamma)\beta}c\epsilon\sum_{k=1}^m m^{(s-\gamma)\beta}\le 2^{(s-\gamma)\beta}c\epsilon m^{(s-\gamma)\beta+1}=2^{(s-\gamma)\beta}cC_0,\end{aligned} \tag{92}$$

Where c in (92) only depends on the constants in Assumption 1. So if C_0 is small such that

$$2^{(s-\gamma)\beta}cC_0\le R, \tag{93}$$

then we have $f_i\in B^s(R),i=1,2,\cdots,M$.

Using Lemma 18, we have

$$\begin{aligned}\mathrm{KL}\left(\rho_{f_i}^n,\rho_{f_0}^n\right)&=\frac{n}{2\bar{\sigma}^2}\|f_i\|_{L^2(\mathcal{X},\mu)}^2\\&=\frac{n\epsilon}{2\bar{\sigma}^2}\sum_{k=1}^m\lambda_{m+k}^{\gamma}\left(\omega_k^{(i)}\right)^2\\&\le\frac{n\epsilon Cm^{-\gamma\beta+1}}{2\bar{\sigma}^2}=\frac{n}{2\bar{\sigma}^2}CC_0m^{-s\beta},\end{aligned}$$

Where C only depends on the constants in Assumption 1. Recall that $M\ge 2^{m/8}$ implies $\ln M\ge\frac{\ln 2}{8}m$. For a fixed $a\in(0,\frac{1}{8})$, since $m=n^{\frac{1}{s\beta+1}}$, letting

$$\mathrm{KL}\left(\rho_{f_i}^n,\rho_{f_0}^n\right)\le\frac{n}{2\bar{\sigma}^2}CC_0m^{-s\beta}\le a\frac{\ln 2}{8}m\le a\ln M, \tag{94}$$

we have

$$C_0 \le \frac{\bar{\sigma}^2 \ln 2}{4C} a. \tag{95}$$

So we can choose $C_0 = c'a$ such that (93) and (95) are satisfied, where c' only depends on the constants in Assumption 1.

Denote $\left\{\rho_{f_i}^n, f_i \in B^s(R)\right\}$ as a family of probability distribution index by f_i, then (94) implies the second condition in Lemma 17 holds. Further, using (91), we have

$$d\left(f_i, f_j\right)^2 = \|f_i - f_j\|_{[\mathcal{H}]^\gamma}^2 = \epsilon \sum_{k=1}^{m} \left(\omega_k^{(i)} - \omega_k^{(j)}\right)^2 \ge \frac{\epsilon m}{8} = \frac{c'a}{8} m^{-(s-\gamma)\beta} \ge c' a n^{-\frac{(s-\gamma)\beta}{s\beta+1}}, \tag{96}$$

where c' only depends on the constants in Assumption 1.

Applying Lemma 17 to (94) and (96), we have

$$\inf_{\hat{f}_n} \sup_{f \in B^s(R)} \mathbb{P}_{\rho_f} \left\{ \left\| \hat{f}_n - f \right\|_{[\mathcal{H}]^\gamma}^2 \ge c' a n^{-\frac{(s-\gamma)\beta}{s\beta+1}} \right\} \ge \frac{\sqrt{M}}{1+\sqrt{M}} \left(1 - 2a - \sqrt{\frac{2a}{\ln M}}\right). \tag{97}$$

When n is sufficiently large so that M is sufficiently large, the probability in the R.H.S. of (97) is larger than $1 - 3a$. For $\delta \in (0,1)$, choose $a = \frac{\delta}{3}$, without loss of generality we assume $a \in (0, \frac{1}{8})$. Then (97) shows that there exists a constant C only depends on the constants in Assumption 1, for all estimator $\hat{f}$, we can find a function $f \in B^s(R)$ and the corresponding distribution $\rho_f \in \mathcal{P}$ such that, with probability at least $1 - \delta$,

$$\left\| \hat{f} - f \right\|_{[\mathcal{H}]^\gamma}^2 \ge C \delta n^{-\frac{(s-\gamma)\beta}{s\beta+1}}.$$

So we finish the proof. (In fact, it can be argued that the constant C only depends on the constants in 1, in dependent of s).

7.5 Shift-invariant kernels

Let μ be the uniform measure on $[-\pi, \pi)^d$. It is well known that the Fourier basis

$$\phi_{\boldsymbol{m}}(x) \coloneqq \exp(i\langle \boldsymbol{m}, x\rangle)$$

are orthonormal in $L^2([-\pi,\pi)^d, \mu)$:

$$\int_{[-\pi,\pi)^d} \phi_{\boldsymbol{m}}(x)\phi_{\boldsymbol{m}'}(x)\mathrm{d}\mu(x) = \frac{1}{(2\pi)^d} \int_{[-\pi,\pi)^d} \phi_{\boldsymbol{m}}(x)\phi_{\boldsymbol{m}'}(x)\mathrm{d}x = \mathbf{1}_{\{\boldsymbol{m}=\boldsymbol{m}'\}}.$$

Now suppose k is a kernel on $[-\pi,\pi)^d$ satisfying

$$k(x,y) = g\big((x-y) \bmod [-\pi,\pi)^d\big).$$

Then, noticing that $\phi_{\boldsymbol{m}}(x)$ is periodic, we have

$$\begin{aligned}
\int_{[-\pi,\pi)^d} k(x,y)\phi_{\boldsymbol{m}}(x)\mathrm{d}\mu(x) &= \int_{[-\pi,\pi)^d} g\big((x-y) \bmod [-\pi,\pi)^d\big) \exp(i\langle \boldsymbol{m}, x\rangle)\mathrm{d}\mu(x) \\
&= \int_{[-\pi,\pi)^d} g\big(z\big) \exp(i\langle \boldsymbol{m}, y+z\rangle)\mathrm{d}\mu(z) \\
&= \exp(i\langle \boldsymbol{m}, y\rangle) \int_{[-\pi,\pi)^d} g\big(z\big) \exp(i\langle \boldsymbol{m}, z\rangle)\mathrm{d}\mu(z).
\end{aligned}$$

It shows that $\phi_{\boldsymbol{m}}(x)$ is an eigenfunction of the integral operator T associated with k. Since $|\phi_{\boldsymbol{m}}(x)| \leq 1$, that is, the eigenfunctions are uniformly bounded, we conclude that the embedding index $\alpha_0 = \frac{1}{\beta}$.

7.6 Spherical harmonics and dot-product kernels

Let us consider the unit d-sphere $\mathbb{S}^d = \{x \in \mathbb{R}^{d+1} \mid \|x\| = 1\}$ and denote by σ the uniform measure on $\mathbb{S}^d$. The eigen-system of spherical Laplacian $\Delta_{\mathbb{S}^d}$ yields an orthogonal decomposition

$$L^2(\mathbb{S}^d, \sigma) = \bigoplus_{n=0}^{\infty} \mathcal{H}_n(\mathbb{S}^d),$$

where $\mathcal{H}_n(\mathbb{S}^d)$ is the subspace of homogenenous harmonic polynomials of degree n and each $Y_n \in \mathcal{H}_n(\mathbb{S}^d)$ is an eigenfunction of $\Delta_{\mathbb{S}^d}$ corresponding to eigenvalue $-n(n+d-1)$. In particular, we can take an orthonormal basis

$$\{Y_{n,l}, l = 1, \ldots, a_n,\ n = 0, 1, \ldots\},$$

where $Y_{n,l} \in \mathcal{H}_n(\mathbb{S}^d)$ and

$$a_n \coloneqq \dim \mathcal{H}_n(\mathbb{S}^d) = \binom{n+d}{n} - \binom{n-2+d}{n-2}.$$

Such an orthonormal basis is often referred to as the *spherical harmonics*. Although the specific choice of $Y_{n,l}$ can vary, the sum

$$Z_n(x, y) = \sum_{l=1}^{a_n} Y_{n,l}(x) Y_{n,l}(y)$$

is invariant. Moreover, $Z_n(x, y)$ depends only on $\langle x, y\rangle$ and satisfies (Dai and Xu, 2013, Corollary 1.2.7)

$$|Z_n(x, y)| \leq Z_n(x, x) = a_n, \quad \forall x, y \in \mathbb{S}^d.$$

The following Funk-Hecke formula is important, see also Dai and Xu (2013, Theorem 1.2.9).

Theorem 20 (Funk-Hecke formula) *Let $d \geq 3$ and f be an integrable function such that $\int_{-1}^{1} |f(t)|(1-t^2)^{d/2-1}\mathrm{d}t$ is finite. Then for every $Y_n \in \mathcal{H}_n(\mathbb{S}^d)$,*

$$\frac{1}{\omega_d} \int_{\mathbb{S}^d} f(\langle x, y\rangle) Y_n(y) \mathrm{d}\sigma(y) = \mu_k(f) Y_n(x), \quad \forall x \in \mathbb{S}^d, \tag{98}$$

where $\mu_n(f)$ is a constant defined by

$$\mu_n(f) = \omega_d \int_{-1}^{1} f(t) \frac{C_n^{\lambda}(t)}{C_n^{\lambda}(1)} (1-t^2)^{\frac{d-2}{2}} \mathrm{d}t,$$

and ω_d is the surface area of $\mathbb{S}^d$.

Suppose k is a dot-product kernel. Recalling the definition of the integral operator T associated with k, (98) shows that elements in $\mathcal{H}_n(\mathbb{S}^d)$, in particular $Y_{n,l}$, are eigenfunctions of T. Therefore, we obtain the following Mercer's decomposition:

$$k(x,y) = \sum_{n=0}^{\infty} \mu_n \sum_{l=1}^{a_n} Y_{n,l}(x) Y_{n,l}(y). \tag{99}$$

Proposition 21 *Let k be an dot-product kernel satisfying $\mu_n \asymp n^{-d\beta}$ for some $\beta > 1$, where μ_n is defined in* (99)*. Then, the EDR of the corresponding RKHS is β and the embedding index $\alpha_0 = \beta^{-1}$.*

Proof Notice that μ_n is an eigenvalue of multiplicity a_n. Then, the eigenvalue decay rate is easily obtained by the estimation $a_n \asymp n^{d-1}$ and $\sum_{r=0}^{n} a_r \asymp n^d$. Considering the equivalent definition of the embedding property (13), we have

$$\begin{aligned}
\sum_{n=0}^{\infty} \mu_n^{\alpha} \sum_{l=1}^{a_n} Y_{n,l}(x)^2 = \sum_{n=0}^{\infty} \mu_n^{\alpha} Z_n(x,x) &\leq \sum_{n=0}^{\infty} a_n \mu_n^{\alpha} \\
&\leq \sum_{n=0}^{\infty} C n^{d-1} n^{-\alpha d \beta} = C \sum_{n=0}^{\infty} n^{-1-d(\alpha\beta-1)} \\
&< \infty \quad \text{if} \quad \alpha > \frac{1}{\beta}.
\end{aligned}$$

■

Acknowledgments and Disclosure of Funding

The corresponding author was supported in part by the National Natural Science Foundation of China (Grant 11971257), Beijing Natural Science Foundation (Grant Z190001), National Key R&D Program of China (2020AAA0105200), and Beijing Academy of Artificial Intelligence.

Appendix A.

In this appendix, we introduce some useful results of real interpolation and Lorentz spaces (Tartar, 2007, Chapter 22-26).

A.1 Real interpolation and the Reiteration theorem

We first introduce the definition of real interpolation through the K-method. For two normed spaces $E_i, i = 0, 1$, denote their norms as $\|\cdot\|_i, i = 0, 1$.

Definition 22 (K-functional) *Let E_0 and E_1 be two normed spaces, continuously embedded into a topological vector space $\mathcal{E}$ ((E_0, E_1) is a compatible couple). For $a \in E_0 + E_1$ and $t > 0$, define the K-functional by*

$$K(t; a) = \inf_{a = a_0 + a_1} \left(\|a_0\|_0 + t \|a_1\|_1 \right).$$

Definition 23 (Real interpolation) *Let E_0 and E_1 be two normed spaces, continuously embedded into a topological vector space $\mathcal{E}$ ((E_0, E_1) is a compatible couple). For $0 < \theta < 1$ and $1 \le p \le \infty$ (or for $\theta = 0, 1$ with $p = \infty$), the real interpolation space is defined by*

$$(E_0, E_1)_{\theta,p} = \left\{ a \in E_0 + E_1 \mid t^{-\theta} K(t; a) \in L^p\left(\mathbb{R}^+; \frac{\mathrm{d}t}{t}\right) \right\},$$

with the norm

$$\|a\|_{(E_0,E_1)_{\theta,p}} = \left\| t^{-\theta} K(t; a) \right\|_{L^p(\mathbb{R}^+; \mathrm{d}t/t)}.$$

Lemma 24 *If $0 < \theta < 1$ and $1 \le p \le q \le \infty$, we have*

$$(E_0, E_1)_{\theta,p} \subset (E_0, E_1)_{\theta,q}, \quad \text{with continuous embedding.}$$

The following lemma gives the result of exchanging the two spaces E_0, E_1.

Lemma 25 *One has $(E_1, E_0)_{\theta,p} = (E_0, E_1)_{1-\theta,p}$ for $0 < \theta < 1$ and $1 \le p \le \infty$; the same result holds for $\theta = 0$ or 1, and $p = 1$ or $p = \infty$.*

The following Lions–Peetre Reiteration Theorem is an important property of real interpolation spaces.

Theorem 26 (Reiteration theorem) *If $0 \le \theta_0 \ne \theta_1 \le 1$, and the two normed spaces F_0, F_1 satisfy that*

$$(E_0, E_1)_{\theta_0,1} \subset F_0 \subset (E_0, E_1)_{\theta_0,\infty};$$
$$(E_0, E_1)_{\theta_1,1} \subset F_1 \subset (E_0, E_1)_{\theta_1,\infty}.$$

Then for $0 < \theta < 1$ and $1 \le p \le \infty$, denote $\eta = (1 - \theta)\theta_0 + \theta\theta_1$, we have

$$(F_0, F_1)_{\theta,p} = (E_0, E_1)_{\eta,p}, \quad \text{with equivalent norms.}$$

Remark 27 *This theorem implies that, if we replace F_0 with any space $\widetilde{F_0}$ satisfying $(E_0, E_1)_{\theta_0,1} \subset \widetilde{F_0} \subset (E_0, E_1)_{\theta_0,\infty}$, the real interpolation space remains 'unchanged', i.e., $(F_0, F_1)_{\theta,p} \cong \left(\widetilde{F_0}, F_1\right)_{\theta,p}$.*

A.2 Lorentz space

Definition 28 (Lorentz space) *For $1 < p < \infty$ and $1 \le q \le \infty$, the Lorentz space $L^{p,q}(\mathcal{X}, \mu)$ is defined as*

$$L^{p,q}(\mathcal{X}, \mu) = \left(L^1(\mathcal{X}, \mu), L^\infty(\mathcal{X}, \mu)\right)_{\frac{1}{p'},q},$$

where $\frac{1}{p'} + \frac{1}{p} = 1$.

Using Lemma 24, it is easy to show that $L^{p,q_1}(\mathcal{X}, \mu) \subseteq L^{p,q_2}(\mathcal{X}, \mu)$ for $1 \le q_1 \le q_2 \le \infty$. In addition, the following lemma gives the relation between Lorentz space and L^p space.

Lemma 29 *For $1 < p < \infty$, 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.

Appendix B. Auxiliary results

Lemma 30 *For any $\lambda > 0$ and $s \in [0, 1]$, we have*

$$\sup_{t \geq 0} \frac{t^s}{t + \lambda} \leq \lambda^{s-1}.$$

Proof Since $a^s \leq a + 1$ for any $a \geq 0$ and $s \in [0, 1]$, the lemma follows from

$$\left(\frac{t}{\lambda}\right)^s \leq \frac{t}{\lambda} + 1 = \frac{t + \lambda}{\lambda}.$$

■

Lemma 31 *If $\lambda_i \asymp i^{-\beta}$, we have*

$$\mathcal{N}(\nu) \asymp \nu^{\frac{1}{\beta}}.$$

Proof Since $c\, i^{-\beta} \leq \lambda_i \leq C i^{-\beta}$, we have

$$\begin{aligned}\mathcal{N}(\nu) &= \sum_{i=1}^{\infty} \frac{\lambda_i}{\lambda_i + \nu^{-1}} \leq \sum_{i=1}^{\infty} \frac{C i^{-\beta}}{C i^{-\beta} + \nu^{-1}} = \sum_{i=1}^{\infty} \frac{C}{C + \nu^{-1} i^{\beta}} \\ &\leq \int_0^{\infty} \frac{C}{\nu^{-1} x^{\beta} + C} \mathrm{d}x = \nu^{\frac{1}{\beta}} \int_0^{\infty} \frac{C}{y^{\beta} + C} \mathrm{d}y \leq C_1 \nu^{\frac{1}{\beta}}.\end{aligned}$$

for some constant C_1. Similarly, we can prove

$$\mathcal{N}(\nu) \geq C_2 \nu^{\frac{1}{\beta}},$$

for some constant C_2. ■

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 32 *Let $(\mathcal{X}, \mathcal{B}, \mu)$ be a probability space, $\mathcal{H}$ be a separable Hilbert space. Suppose that $A_1, \cdots, A_n$ 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. $x \in \mathcal{X}$, and there exists a self-adjoint positive semi-definite trace class operator V with $\mathbb{E}A_i^2 \preceq V$. Then for $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have*

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

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

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

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

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

Lemma 34 (Cordes inequality) *Let A and B be two positive bounded linear operators on a separable Hilbert space. Then we have*

$$\|A^s B^s\| \leq \|AB\|^s, \qquad \text{when } 0 \leq s \leq 1.$$

The following lemma is a corollary of Lin et al. (2018, Lemma 5.8).

Lemma 35 *Suppose that A and B are two positive self-adjoint operators on some Hilbert space, then*

- *for $r \in (0,1]$, we have*
$$\|A^r - B^r\| \leq \|A - B\|^r.$$
- *for $r \geq 1$, denote $c = \max(\|A\|, \|B\|)$, we have*
$$\|A^r - B^r\| \leq r c^{r-1} \|A - B\|.$$

Appendix C. Details of experiments

First, we prove that the series in (17) converges and $f^*(x)$ is continuous on $(0,1)$ for $0 < s < \frac{1}{\beta} = 0.5$. We begin with the computation of the sum of first N terms of $\{\sin 2k\pi x + \cos 2k\pi x\}$, note that

$$\begin{aligned}
&-2\sin(\pi x)\left(\sin(2\pi x) + \sin(4\pi x) + \cdots + \sin(2N\pi x)\right) \\
&= \left[\cos(2\pi+\pi)x - \cos(2\pi-\pi)x\right] + \left[\cos(4\pi+\pi)x - \cos(4\pi-\pi)x\right] \\
&\quad + \cdots + \left[\cos(2N\pi+\pi)x - \cos(2N\pi-\pi)x\right] \\
&= \cos(2N\pi+\pi)x - \cos\pi x.
\end{aligned}$$

So we have

$$\left|\left(\sin(2\pi x) + \sin(4\pi x) + \cdots + \sin(2N\pi x)\right)\right| = \frac{\left|\cos(2N\pi+\pi)x - \cos\pi x\right|}{\left|2\sin(\pi x)\right|}; \tag{100}$$

Similarly, we have

$$\left|\left(\cos(2\pi x) + \cos(4\pi x) + \cdots + \cos(2N\pi x)\right)\right| = \frac{\left|\sin(2N\pi+\pi)x - \sin\pi x\right|}{\left|2\sin(\pi x)\right|}. \tag{101}$$

Note that (100) and (101) are uniformly bounded in $[\delta_0, 1-\delta_0]$ for any $\delta_0 > 0$ and N. In addition, $\{k^{-(s+0.5)}\}$ is monotone and decreases to zero. Use the Dirichlet criterion and we know that the series in (17) is uniformly convergence in $[\delta_0, 1-\delta_0]$. Due to the arbitrariness of δ_0, we know that the series converges and $f^*(x)$ is continuous on $(0,1)$.

Next, we prove that the series in (18) converges and $f^*(x)$ is continuous on $(0,1)$ for $0 < s < \frac{1}{\beta} = 0.5$. We begin with the computation of the sum of first N terms of $e_{2k-1}(x)$,

$$\begin{aligned}
&-2\sin(\pi x)\left(\sin\left(\frac{\pi x}{2}\right)+\sin\left(\frac{5\pi x}{2}\right)+\cdots+\sin\left(\frac{(4N-3)\,\pi x}{2}\right)\right)\\
&=\left[\cos\left(\pi+\frac{\pi}{2}\right)x-\cos\left(\pi-\frac{\pi}{2}\right)x\right]+\left[\cos\left(\frac{5\pi}{2}+\pi\right)x-\cos\left(\frac{5\pi}{2}-\pi\right)x\right]\\
&\quad+\cdots+\left[\cos\left(\frac{(4N-3)\,\pi}{2}+\pi\right)x-\cos\left(\frac{(4N-3)\,\pi}{2}-\pi\right)x\right]\\
&=\cos\left(\frac{(4N-1)\,\pi}{2}\right)x-\cos\frac{\pi}{2}x.
\end{aligned}$$

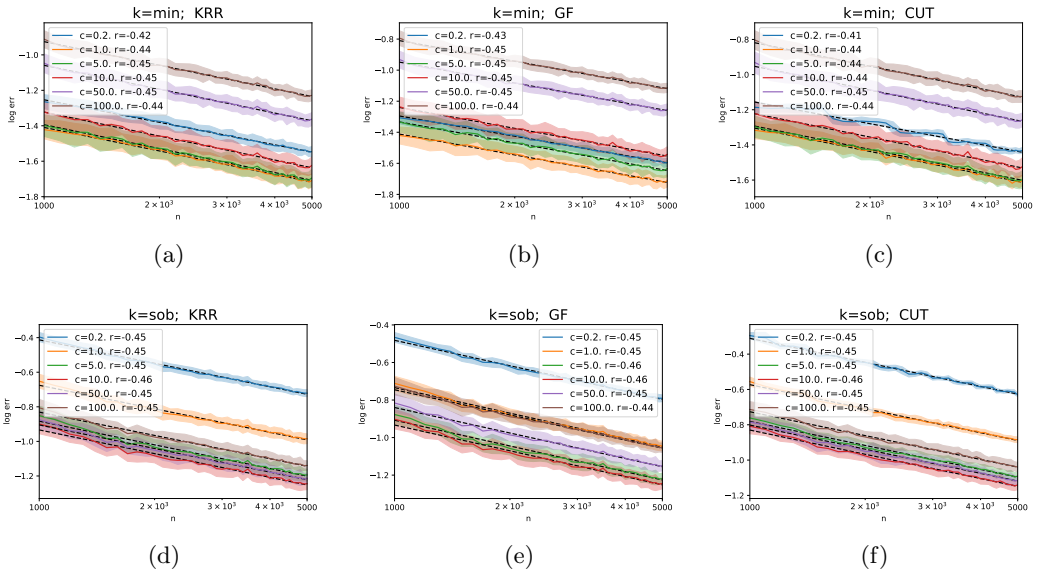
So we have

$$\left|\sin\left(\frac{\pi x}{2}\right)+\sin\left(\frac{5\pi x}{2}\right)+\cdots+\sin\left(\frac{(4N-3)\,\pi x}{2}\right)\right|=\frac{\left|\cos\left(\frac{(4N-1)\pi}{2}\right)x-\cos\frac{\pi}{2}x\right|}{|2\sin(\pi x)|},$$

which is uniformly bounded in $[\delta_0, 1-\delta_0]$ for any $\delta_0 > 0$ and N.

Note that $\{k^{-(s+0.5)}\}$ is monotone and decreases to zero. Use the Dirichlet criterion and we know that the series in (18) is uniformly convergence in $[\delta_0, 1-\delta_0]$. Due to the arbitrariness of δ_0, we know that the series converges and $f^*(x)$ is continuous on $(0,1)$.

In Figure 2, we present the results of different choices of c for $\nu = cn^{\frac{\beta}{s\beta+1}}$ in the experiment of Section 5.



(a) (b) (c)

(d) (e) (f)

Figure 2: Error decay curves of two kinds of RKHSs and three kinds of spectral algorithms with different choices of c. Both axes are logarithmic.

References

R. Adams. *Sobolev Spaces. Adams*. Pure and applied mathematics. Academic Press, 1975. URL `https://books.google.co.uk/books?id=JxzpSAAACAAJ`.

R. A. Adams and J. J. Fournier. *Sobolev Spaces*. Elsevier, 2003.

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

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. Kernel ridgeless regression is inconsistent in low dimensions, June 2022.

G. Blanchard and N. Mücke. Optimal rates for regularization of statistical inverse learning problems. *Foundations of Computational Mathematics*, 18:971–1013, 2018.

B. Bordelon, A. Canatar, and C. Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *ICML*, 2020.

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. Celisse and M. Wahl. Analyzing the discrepancy principle for kernelized spectral filter learning algorithms. *J. Mach. Learn. Res.*, 22:76:1–76:59, 2020.

Y. Cho and L. Saul. Kernel methods for deep learning. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009.

F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39:1–49, 2001.

H. Cui, B. Loureiro, F. Krzakala, and L. Zdeborov'a. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. In *NeurIPS*, 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.

L. Dicker, D. P. Foster, and D. J. Hsu. Kernel ridge vs. principal component regression: Minimax bounds and the qualification of regularization operators. *Electronic Journal of Statistics*, 11:1022–1047, 2017.

A. Dieuleveut and F. Bach. Nonparametric stochastic approximation with large step-sizes1. *THE ANNALS*, 44(4):1363–1399, 2016.

D. E. Edmunds and H. Triebel. *Function Spaces, Entropy Numbers, Differential Operators*. Cambridge Tracts in Mathematics. Cambridge University Press, 1996. doi: 10.1017/CBO9780511662201.

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.

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.

Z.-C. Guo, S. Lin, and D.-X. Zhou. Learning theory of distributed spectral algorithms. *Inverse Problems*, 33, 2017.

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.

K.-S. Jun, A. Cutkosky, and F. Orabona. Kernel truncated randomized ridge regression: Optimal rates and low noise acceleration. *ArXiv*, abs/1905.10680, 2019.

M. Kohler and A. Krzyżak. Nonparametric regression estimation using penalized least squares. *IEEE Trans. Inf. Theory*, 47:3054–3059, 2001.

Z. Li, D. Meunier, M. Mollenhauer, and A. Gretton. Optimal rates for regularized conditional mean embedding learning. *ArXiv*, abs/2208.01711, 2022.

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.

S. Lin, X. Guo, and D.-X. Zhou. Distributed learning with regularized least squares. *ArXiv*, abs/1608.03339, 2016.

J. Liu and L. Shi. Statistical optimality of divide and conquer kernel-based functional linear regression. *ArXiv*, abs/2211.10968, 2022.

S. Mendelson and J. Neeman. Regularization in kernel learning. *The Annals of Statistics*, 38(1):526–565, Feb. 2010.

N. Mücke and G. Blanchard. Parallelizing spectrally regularized kernel algorithms. *J. Mach. Learn. Res.*, 19:30:1–30:29, 2018.

L. Pillaud-Vivien, A. Rudi, and F. R. Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. *ArXiv*, abs/1805.10074, 2018.

A. Rastogi and S. Sampath. Optimal rates for the regularized learning algorithms under general source condition. *Frontiers in Applied Mathematics and Statistics*, 3, 2017.

L. Rosasco, E. De Vito, and A. Verri. Spectral methods for regularization in learning theory. *DISI, Universita degli Studi di Genova, Italy, Technical Report DISI-TR-05-18*, 2005.

Y. Sawano. *Theory of Besov spaces*, volume 56. Springer, 2018.

S. Smale and D.-X. Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26:153–172, 2007.

A. Smola, Z. Ovári, and R. C. Williamson. Regularization with dot-product kernels. *Advances in neural information processing systems*, 13, 2000.

S. Spigler, M. Geiger, and M. Wyart. Asymptotic learning curves of kernel methods: empirical data versus teacher–student paradigm. *Journal of Statistical Mechanics: Theory and Experiment*, 2020, 2020.

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.

P. M. Talwai and D. Simchi-Levi. Optimal learning rates for regularized least-squares with a fourier capacity condition. 2022.

L. Tartar. *An introduction to Sobolev spaces and interpolation spaces*, volume 3. Springer Science & Business Media, 2007.

C. Thomas-Agnan. Computing a family of reproducing kernels for statistical applications. *Numerical Algorithms*, 13:21–32, 1996.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York ; London, 1st edition, 2009.

M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

W. Wang and B.-Y. Jing. Gaussian process regression: Optimality, robustness, and relationship with kernel ridge regression. *Journal of Machine Learning Research*, 23(193):1–67, 2022. URL `http://jmlr.org/papers/v23/21-0570.html`.

Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26:289–315, 2007.

Y. Zhang, J. C. Duchi, and M. J. Wainwright. Divide and conquer kernel ridge regression: a distributed algorithm with minimax optimal rates. *J. Mach. Learn. Res.*, 16:3299–3340, 2013.