On the Optimality of Misspecified Kernel Ridge Regression

Haobo Zhang [1] **Yicheng Li** [1] **Weihao Lu** [1] **Qian Lin** [1 2]

Abstract

In the misspecified kernel ridge regression 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 KRR is optimal for all $s \in (0,1)$ is an outstanding problem lasting for years. In this paper, we show that KRR is minimax optimal for any $s \in (0,1)$ when the $\mathcal{H}$ is a Sobolev RKHS.

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 $\mathcal{Y} \subseteq \mathbb{R}$. The regression problem aims to find a function $\hat{f}$ such that the risk

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

is relatively small. It is well known that the conditional mean function given by $f_\rho^*(x) := \mathbb{E}_\rho[\, y \mid x \,] = \int_{\mathcal{Y}} y \mathrm{d}\rho(y|x)$ 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 (generalization error)

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

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 certain structures 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 to a kernel function k defined on $\mathcal{X}$, e.g., Kohler & Krzyżak (2001); Cucker & Smale (2001); Caponnetto & de Vito (2007); Steinwart & Christmann (2008). This paper focuses on the kernel ridge regression (KRR), which constructs an estimator $\hat{f}_\lambda$ by solving the penalized least square problem

$$\hat{f}_\lambda = \arg\min_{f\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^n \left(y_i - f\left(x_i\right)\right)^2 + \lambda\|f\|_{\mathcal{H}}^2\right), \tag{2}$$

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

Since the minimax optimality of KRR has been proved for $f_\rho^* \in [\mathcal{H}]^s, 1 \le s \le 2$ (Caponnetto & de Vito, 2007), a large body of literature has studied the convergence rate of the generalization error of misspecified KRR ($f_\rho^* \notin \mathcal{H}$) and whether the rate is optimal in the minimax sense. It turns out that 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 behavior of the misspecified KRR (see Section 2.2 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 is $n^{-\frac{s\beta}{s\beta+1}}$. The state-of-the-art work Fischer & Steinwart (2020) has already shown that when $\alpha_0 < s \le 2$, the upper bound of the convergence rate of KRR is $n^{-\frac{s\beta}{s\beta+1}}$ and hence is optimal. However, when $f_\rho^* \in [\mathcal{H}]^s$ for some $0 < s \le \alpha_0$, all the existing works need an additional boundedness assumption on 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., $[\mathcal{H}]^s \cap L^\infty(\mathcal{X}, \mu) \subsetneqq [\mathcal{H}]^s$ when $s \le \alpha_0$. Fischer & Steinwart (2020) further reveals that the minimax rate of the excess risk associated to the smaller function space is larger than $n^{-\frac{\alpha\beta}{\alpha\beta+1}}$ for any $\alpha > \alpha_0$. This lower bound of the minimax rate is smaller than the upper bound of the convergence rate and hence they can not prove the minimax optimality of KRR when $s \le \alpha_0$.

It has been an outstanding problem for years whether KRR is minimax optimal for all the $s \in (0,1)$ (Pillaud-Vivien et al., 2018; Fischer & Steinwart, 2020; Liu & Shi, 2022). This paper concludes that, for Sobolev RKHS, KRR is optimal

[1]Center for Statistical Science, Department of Industrial Engineering, Tsinghua University [2]Beijing Academy of Artificial Intelligence, Beijing, China. Correspondence to: Qian Lin <qianlin@tsinghua.edu.cn>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

for all $0 < s < 1$. Thus, we know that KRR is optimal for Sobolev RKHS and all $0 < s \leq 2$. Together with a recent work on the saturation effect where KRR can not be optimal for $s > 2$ (Li et al., 2023), the optimality of KRR for Sobolev RKHS is well understood.

1.1. Related work

Kernel ridge regression has been studied as a special kind of spectral regularization algorithm (Rosasco et al., 2005; Caponnetto, 2006; Bauer et al., 2007; Gerfo et al., 2008; Mendelson & Neeman, 2010). In large part of the literature, the 'hardness' of the regression problem is determined by two parameters: 1. the source condition (s), which characterizes a function's relative 'smoothness' with respect to the RKHS; 2. the eigenvalue decay rate (β) (or capacity, effective dimension equivalently), which characterizes the RKHS itself. These two parameters divide the convergence behavior of KRR or spectral regularization algorithm into different regimes and lead to different convergence rates (Dicker et al., 2017; Blanchard & Mücke, 2018; Lin et al., 2018; Lin & Cevher, 2020; Li et al., 2023, etc.). Caponnetto & de Vito (2007) first proves the optimality when $1 \leq s \leq 2$ and Lin et al. (2018) extends the desired upper bound of the convergence rate to the regime $s + \frac{1}{\beta} > 1$.

KRR in the misspecified case ($f_\rho^* \notin \mathcal{H}$ or $s < 1$) has also been discussed by another important line of work which considers the embedding index (α_0) of the RKHS and performs refined analysis (Steinwart et al., 2009; Dicker et al., 2017; Pillaud-Vivien et al., 2018; Fischer & Steinwart, 2020; Celisse & Wahl, 2020; Li et al., 2022). The desired upper bound $n^{-\frac{s\beta}{s\beta+1}}$ is extended to the regime $s + \frac{1}{\beta} > \alpha_0$, and the minimax optimality is extended to the regime $s > \alpha_0$. It is worth pointing out that 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., $\|f_\rho^*\|_{L^\infty(\mathcal{X},\mu)} \leq B_\infty < \infty$ (Lin & Cevher, 2020; Fischer & Steinwart, 2020; Talwai & Simchi-Levi, 2022; Li et al., 2022, etc). As discussed in the introduction, this will lead to the suboptimality in the $s \leq \alpha_0$ regime.

This paper follows the line of work that considers the embedding index, refines the proof by handling the additional boundedness assumption, and solves the optimality problem for Sobolev RKHS and all $s \in (0, 1)$. In addition, our technical improvement also sheds light on the optimality for general RKHS. Specifically, we replace the boundedness assumption with a far weaker L^q-integrability assumption, which turns out to be reasonable for many RKHS. Note that our results focus on the most frequently used L^2 convergence rate for KRR, and it can be easily extended to $[\mathcal{H}]^\gamma, \gamma \geq 0$ convergence rate (e.g., Fischer & Steinwart 2020) and general spectral regularization algorithms (e.g., Lin et al. 2018).

2. Preliminaries

2.1. Basic concepts

Let $\mathcal{X} \subseteq \mathbb{R}^d$ 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. Then the generalization error (1) can be written as

$$\left\| \hat{f} - f_\rho^* \right\|_{L^2}^2 .$$

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

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

Define the integral operator $T : L^2(\mathcal{X}, \mu) \to L^2(\mathcal{X}, \mu)$ as

$$(Tf)(x) := \int_{\mathcal{X}} k(x, y) f(y) \mathrm{d}\mu(y). \tag{3}$$

It is well known that T is a positive, self-adjoint, trace-class, and hence a compact operator (Steinwart & Scovel, 2012). The spectral theorem for self-adjoint compact operators and Mercer's decomposition theorem yield that

$$k(x, y) = \sum_{i\in N} \lambda_i e_i(x) e_i(y),$$
$$T = \sum_{i\in N} \lambda_i \langle \cdot, e_i \rangle_{L^2} e_i,$$

where N is an at most countable set, the eigenvalues $\{\lambda_i\}_{i\in N} \subseteq (0, \infty)$ is a non-increasing summable sequence, and $\{e_i\}_{i\in N}$ are the corresponding eigenfunctions. Denote the samples as $X = (x_1, \cdots, x_n)$ and $\mathbf{y} = (y_1, \cdots, y_n)'$. The representer theorem (see, e.g., Steinwart & Christmann 2008) gives an explicit expression of the KRR estimator defined by (2), i.e.,

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

where

$$\mathbb{K}(X, X) = (k(x_i, x_j))_{n\times n},$$

and

$$\mathbb{K}(x, X) = (k(x, x_1), \cdots, k(x, x_n)).$$

We also need to introduce the interpolation spaces of RKHS. For any $s \geq 0$, the fractional power integral operator $T^s : L^2(\mathcal{X}, \mu) \to L^2(\mathcal{X}, \mu)$ is defined as

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

and the interpolation space $[\mathcal{H}]^s, s \geq 0$ of $\mathcal{H}$ is defined as

$$[\mathcal{H}]^s := \operatorname{Ran} T^{\frac{s}{2}} = \left\{ \sum_{i \in N} \lambda_i^{\frac{s}{2}} a_i e_i \mid \sum_{i \in N} a_i^2 < \infty \right\}, \quad (4)$$

with the inner product

$$\langle f, g \rangle_{[\mathcal{H}]^s} = \left\langle T^{-\frac{s}{2}} f, T^{-\frac{s}{2}} g \right\rangle_{L^2}. \quad (5)$$

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

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 3.1 for detailed discussions.

2.2. Assumptions

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

Assumption 2.1 (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

$$c i^{-\beta} \leq \lambda_i \leq C i^{-\beta}, \quad \forall i \in N.$$

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.

Assumption 2.2 (Embedding index)**.** We say that $[\mathcal{H}]^\alpha$ has the embedding property for some $\alpha \in [\frac{1}{\beta}, 1]$, if there is a constant $0 < A < \infty$ such that

$$\|[\mathcal{H}]^\alpha \hookrightarrow L^\infty(\mathcal{X}, \mu)\| \leq A, \quad (6)$$

where $\|\cdot\|$ denotes the operator norm of the embedding. Then we define the *embedding index* of an RKHS $\mathcal{H}$ as

$$\alpha_0 = \inf \{\alpha : [\mathcal{H}]^\alpha \text{ has the embedding property}\}.$$

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

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

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

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

The larger α is, the weaker the embedding property is. 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 & Steinwart (2020, Lemma 10) also shows that α can not be less than $\frac{1}{\beta}$.

Note that the embedding property (6) holds 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 2.3 (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 2.4 (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 & Cevher, 2020; Fischer & 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 directly implies both Assumption 2.4 and the boundedness of f_ρ^*.

2.3. Review of state-of-the-art results

For the convenience of comparing our results with previous works, we review state-of-the-art upper and lower bounds of the convergence rate of KRR (Fischer & Steinwart, 2020).

Proposition 2.5 (Upper bound)**.** *Suppose that Assumption 2.1,2.2, 2.3 and 2.4 hold for $0 < s \leq 2$ and $\frac{1}{\beta} \leq \alpha_0 < 1$. Furthermore, suppose that there exists a constant $B_\infty > 0$ such that $\|f_\rho^*\|_{L^\infty(\mathcal{X},\mu)} \leq B_\infty$. Let $\hat{f}_\lambda$ be the KRR estimator defined by (2). Then in the case of $s + \frac{1}{\beta} > \alpha_0$, by choosing $\lambda \asymp n^{-\frac{\beta}{s\beta+1}}$, for any fixed $\delta \in (0, 1)$, when n is sufficiently large, with probability at least $1 - \delta$, we have*

$$\left\| \hat{f}_\lambda - f_\rho^* \right\|_{L_2}^2 \leq \left(\ln \frac{4}{\delta} \right)^2 C n^{-\frac{s\beta}{s\beta+1}},$$

where C is a constant independent of n and δ.

Remark 2.6. When $s+\frac{1}{\beta}\le\alpha_0$, Fischer & Steinwart (2020, Theorem 1) also proves the state-of-the-art upper bound of the convergence rate $(n/\log^r(n))^{-\frac{s}{\alpha}}, \forall r>0$ and $\alpha>\alpha_0$ which can be arbitrarily close to α_0. We should note that $s+\frac{1}{\beta}>\alpha_0$ is always satisfied for Sobolev RKHS (see Section 3.1), so we focus on $s+\frac{1}{\beta}>\alpha_0$ in this paper.

Proposition 2.7 (Minimax lower bound)**.** *Let μ be a probability distribution on $\mathcal{X}$ such that Assumption 2.1 and 2.2 are satisfied for $\frac{1}{\beta}\le\alpha_0<1$. Let $\mathcal{P}$ consist of all the distributions on $\mathcal{X}\times\mathcal{Y}$ satisfying 2.3, 2.4 for $0<s\le 2$ and with marginal distribution μ. For a constant $B_\infty>0$, let $\mathcal{P}_\infty$ consists of all the distributions on $\mathcal{X}\times\mathcal{Y}$ such that $\|f_\rho^*\|_{L^\infty(\mathcal{X},\mu)}\le B_\infty$. Then for any $\alpha>\alpha_0$, 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 $\rho\in\mathcal{P}\cap\mathcal{P}_\infty$ such that, with probability at least $1-\delta$, we have*

$$\left\|\hat{f}-f_\rho^*\right\|_{L^2}^2\ge C\delta n^{-\frac{\max\{s,\alpha\}\beta}{\max\{s,\alpha\}\beta+1}}.$$

Remark 2.8. Under the precondition $s+\frac{1}{\beta}>\alpha_0$, (1) when $s>\alpha_0$, since there always exists $\alpha_0<\alpha\le s$, the upper bound of the convergence rate in Proposition 2.5 coincides with the minimax lower bound in Proposition 2.7 and hence is minimax optimal; (2) but when $s\le\alpha_0$, existing results fail to show the optimality of KRR. The same gap between the upper and lower bound exists for gradient descent and stochastic gradient descent with multiple passes (Pillaud-Vivien et al., 2018).

Note that in Proposition 2.7, we consider the distributions in $\mathcal{P}\cap\mathcal{P}_\infty$. If we consider all the distributions in $\mathcal{P}$, we have the following minimax lower bound, which is often referred to as the information-theoretic lower bound (see, e.g., Rastogi & Sampath 2017).

Proposition 2.9 (Information-theoretic lower bound)**.** *Let μ be a probability distribution on $\mathcal{X}$ such that Assumption 2.1 is satisfied. Let $\mathcal{P}$ consist of all the distributions on $\mathcal{X}\times\mathcal{Y}$ satisfying 2.3, 2.4 for $0<s\le 2$ and with marginal distribution μ. Then 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*

$$\left\|\hat{f}-f_\rho^*\right\|_{L^2}^2\ge C\delta n^{-\frac{s\beta}{s\beta+1}}.$$

3. Main Results

The main results of this paper aim to remove the boundedness assumption $\|f_\rho^*\|_{L^\infty(\mathcal{X},\mu)}\le B_\infty$. We state the main theorem in terms of general RKHS, and make detailed discussions for Sobolev space as a particular case.

Theorem 3.1. *Suppose that Assumption 2.1, 2.2, 2.3 and 2.4 hold for $0<s\le 2$ and $\frac{1}{\beta}\le\alpha_0<1$. Suppose that $f_\rho^*\in L^q(\mathcal{X},\mu)$ and $\|f_\rho^*\|_{L^q(\mathcal{X},\mu)}\le C_q<\infty$ for some $q>\frac{2(s\beta+1)}{2+(s-\alpha_0)\beta}$. Let $\hat{f}_\lambda$ be the KRR estimator defined by (2). Then, in the case of $s+\frac{1}{\beta}>\alpha_0$, by choosing $\lambda\asymp n^{-\frac{\beta}{s\beta+1}}$, for any fixed $\delta\in(0,1)$, when n is sufficiently large, with probability at least $1-\delta$, we have*

$$\left\|\hat{f}_\lambda-f_\rho^*\right\|_{L^2}^2\le\left(\ln\frac{4}{\delta}\right)^2 Cn^{-\frac{s\beta}{s\beta+1}},$$

where C is a constant independent of n and δ.

Remark 3.2. In Theorem 3.1, we replace the boundedness assumption with a L^q-integrability assumption and prove the same upper bound of the convergence rate as Proposition 2.5 in the case of $s+\frac{1}{\beta}>\alpha_0$. As shown in the following, both $s+\frac{1}{\beta}>\alpha_0$ and $q>\frac{2(s\beta+1)}{2+(s-\alpha_0)\beta}$ are naturally satisfied for Sobolev RKHS.

Remark 3.3. RKHS with uniformly bounded eigenfunctions, i.e., $\sup_{i\in N}\|e_i\|_{L^\infty}<\infty$ are frequently considered (Mendelson & Neeman, 2010; Steinwart et al., 2009; Pillaud-Vivien et al., 2018). For this kind of RKHS, the Assumption 2.2 holds for $\alpha_0=\frac{1}{\beta}$ (Fischer & Steinwart, 2020, Lemma 10), hence $s+\frac{1}{\beta}>\alpha_0$ is satisfied. In addition, the assumption $q>\frac{2(s\beta+1)}{2+(s-\alpha_0)\beta}$ in Theorem 3.1 turns into $q>2$. Recalling that $[\mathcal{H}]^0\subseteq L^2(\mathcal{X},\mu)$ when $s=0$, so it is reasonable to assume that the functions in $[\mathcal{H}]^s, s>0$, is L^q-integrable for some $q>2$.

Remark 3.4. If the RKHS $\mathcal{H}$ satisfies that $[\mathcal{H}]^s\hookrightarrow L^q$, i.e., $[\mathcal{H}]^s\cap L^q=[\mathcal{H}]^s$ for some q satisfying the integrability required in Theorem 3.1. Using Proposition 2.9, the minimax lower bound will be still $n^{-\frac{s\beta}{s\beta+1}}$ even when making the assumption $\|f_\rho^*\|_{L^q}<\infty$.

3.1. Optimality for Sobolev RKHS

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

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

where $m:=\min\{k\in\mathbb{N}:k>r\}$. (For details of real interpolation of Banach spaces, we refer to Sawano (2018,

Chapter 4.2.2)). 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 & Triebel 1996)

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

Furthermore, for the interpolation space of $H^r(\mathcal{X})$ under Lebesgue measure defined by (4), we have (see, e.g., Steinwart & Scovel 2012, Theorem 4.6), for $s > 0$,

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

Now we begin to introduce the embedding theorem of (fractional) Sobolev space from 7.57 of Adams (1975), which is crucial when applying Theorem 3.1 to Sobolev RKHS.

Proposition 3.5 (Embedding theorem)**.** *Let $H^r(\mathcal{X}), r > 0$ be defined as above, we have*

(i) if $d > 2r$, then $H^r(\mathcal{X}) \hookrightarrow L^q(\mathcal{X})$ for $2 \le q \le \frac{2d}{d-2r}$.

(ii) if $d = 2r$, then $H^r(\mathcal{X}) \hookrightarrow L^q(\mathcal{X})$ for $2 \le q < \infty$.

(iii) if $d < 2(r-j)$ for some nonnegative integer j, then $H^r(\mathcal{X}) \hookrightarrow C^{j,\gamma}(\mathcal{X}), \gamma = r - j - \frac{d}{2}$,

where $C^{j,\gamma}(\mathcal{X})$ denotes the Hölder space and $\hookrightarrow$ denotes the continuous embedding.

On the one hand, (iii) of Proposition 3.5 shows that, for a Sobolev RKHS $\mathcal{H} = H^r(\mathcal{X}), r > \frac{d}{2}$ and any $\alpha > \frac{1}{\beta} = \frac{d}{2r}$,

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

where $\gamma > 0$. So the Assumption 2.2 holds for $\alpha_0 = \frac{1}{\beta}$, and thus $\frac{2(s\beta+1)}{2+(s-\alpha_0)\beta} = 2$. On the other hand, (i) of Proposition 3.5 shows that if $d > 2rs$,

$$[H^r(\mathcal{X})]^s = H^{rs}(\mathcal{X}) \hookrightarrow L^q(\mathcal{X}),$$

where $q = \frac{2d}{d-2rs} = \frac{2}{1-s\beta} > 2$. In addition, (ii) and (iii) of Proposition 3.5 show that $q > 2$ also holds if $d \le 2rs$. Therefore, for any $0 < s \le 2$, we have

$$\text{Assumption 2.2; } s + \frac{1}{\beta} > \alpha_0; \ q > \frac{2(s\beta+1)}{2+(s-\alpha_0)\beta} = 2$$

hold simultaneously.

Now we are ready to state a theorem as the corollary of Theorem 3.1 and Proposition 3.5, which implies the optimality of KRR for Sobolev RKHS and any $0 < s \le 2$.

Theorem 3.6. *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a bounded domain with a smooth boundary. The RKHS is $\mathcal{H} = H^r(\mathcal{X})$ for some $r > d/2$. Suppose that the distribution ρ satisfies that $\|f_\rho^*\|_{[\mathcal{H}]^s} \le R$ for $0 < s \le 2$, the noise satisfies Assumption 2.4, and the marginal distribution μ on $\mathcal{X}$ has Lebesgue density $0 < c \le p(x) \le C$ for two constants c and C. Let $\hat{f}_\lambda$ be the KRR estimator defined by (2). Then, by choosing $\lambda \asymp n^{-\frac{\beta}{s\beta+1}}$, for any fixed $\delta \in (0,1)$, when n is sufficiently large, with probability at least $1-\delta$, we have*

$$\left\|\hat{f}_\lambda - f_\rho^*\right\|_{L^2}^2 \le \left(\ln\frac{4}{\delta}\right)^2 C n^{-\frac{s\beta}{s\beta+1}},$$

where C is a constant independent of n and δ.

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 $c \le \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \le C$.

Remark 3.7 (Optimality for Sobolev RKHS). Without the boundedness assumption, Theorem 3.6 proves the same upper bound of the convergence rate $n^{-\frac{s\beta}{s\beta+1}}$. Together with the information-theoretic lower bound in Proposition 2.9, we prove the optimality of KRR for all $0 < s \le 2$, while state-of-the-art result Fischer & Steinwart (2020, Corollary 5) can only prove for $\frac{1}{\beta} < s \le 2$. Since when $s > 2$, the saturation effect of KRR has been proved in a recent work Li et al. (2023), the optimality of KRR for Sobolev spaces is well understood.

3.2. Sketch of proof

In this subsection, we present the sketch of the proof of Theorem 3.1. The rigorous proof of Theorem 3.1, Proposition 2.9, and Theorem 3.6 will be in the appendix. We refer to Fischer & Steinwart (2020, Chapter 6) for the proof of Proposition 2.5 and 2.7.

Our proofs are based on the standard integral operator techniques dating back to Smale & Zhou (2007), and we refine the proof to handle the unbounded case. Let us first introduce some frequently used notations. Define the sample operator as

$$K_x : \mathbb{R} \to \mathcal{H}, \ \ y \mapsto yk(x, \cdot),$$

and its adjoint operator

$$K_x^* : \mathcal{H} \to \mathbb{R}, \ \ f \mapsto f(x).$$

Next we define the sample covariance operator $T_X : \mathcal{H} \to \mathcal{H}$ as

$$T_X := \frac{1}{n}\sum_{i=1}^n K_{x_i} K_{x_i}^*, \tag{7}$$

and the sample basis function

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

Then the KRR estimator (2) can be expressed by (see, e.g., Caponnetto & de Vito (2007, Chapter 5))

$$\hat{f}_\lambda = (T_X + \lambda)^{-1} g_Z.$$

Define f_λ as the unique minimizer given by

$$f_\lambda = \arg\min_{f \in \mathcal{H}} \left(\int_{\mathcal{X}\times\mathcal{Y}} (f(x) - y)^2 \, \mathrm{d}\rho(x, y) + \lambda \|f\|_{\mathcal{H}}^2 \right).$$

Note that the integral operator T can also be seen as a bounded linear operator on $\mathcal{H}$, f_λ can be expressed by

$$f_\lambda = (T + \lambda)^{-1} g, \tag{8}$$

where g is the expectation of g_Z given by

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

The first step in our proof is to decompose the generalization error into two terms, which are often referred to as the approximation error and estimation error,

$$\left\| \hat{f}_\lambda - f_\rho^* \right\|_{L^2} \le \left\| \hat{f}_\lambda - f_\lambda \right\|_{L^2} + \left\| f_\lambda - f_\rho^* \right\|_{L^2}, \tag{9}$$

Then we will show that by choosing $\lambda \asymp n^{-\frac{\beta}{s\beta+1}}$,

$$\left\| f_\lambda - f_\rho^* \right\|_{L^2} \le CR\lambda^{\frac{s}{2}} = CRn^{-\frac{1}{2}\frac{s\beta}{s\beta+1}}; \tag{10}$$

and for any fixed $\delta \in (0, 1)$, when n is sufficiently large, with probability at least $1 - \delta$, we have

$$\left\| \hat{f}_\lambda - f_\lambda \right\|_{L^2} \le \ln \frac{4}{\delta} C \frac{\lambda^{-\frac{1}{2\beta}}}{\sqrt{n}} = \ln \frac{4}{\delta} C n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}}. \tag{11}$$

Plugging (10) and (11) into (9) and we will finish the proof of Theorem 3.1.

The approximation error Suppose that $f_\rho^* = \sum_{i\in N} a_i e_i$. Then using the expression (8) and simple inequality, we will show that

$$\left\| f_\lambda - f_\rho^* \right\|_{L^2}^2 = \sum_{i\in N} \left(\frac{\lambda \lambda_i^{s/2}}{\lambda + \lambda_i} \right)^2 \lambda_i^{-s} a_i^2 \le \lambda^s \|f_\rho^*\|_{[\mathcal{H}]^s}.$$

The estimation error Denote $T_{X\lambda} = T_X + \lambda$ and $T_\lambda = T + \lambda$. We will first rewrite the estimator error as follows

$$\begin{aligned} &\left\| \hat{f}_\lambda - f_\lambda \right\|_{L^2} = \left\| T^{\frac{1}{2}} \left(\hat{f}_\lambda - f_\lambda \right) \right\|_{\mathcal{H}} \\ &\le \left\| T^{\frac{1}{2}} T_\lambda^{-\frac{1}{2}} \right\| \cdot \left\| T_\lambda^{\frac{1}{2}} T_{X\lambda}^{-1} T_\lambda^{\frac{1}{2}} \right\| \cdot \left\| T_\lambda^{-\frac{1}{2}} \left(g_Z - T_{X\lambda} f_\lambda \right) \right\|_{\mathcal{H}}. \end{aligned} \tag{12}$$

For the first term in (12), we have

$$\left\| T^{\frac{1}{2}} T_\lambda^{-\frac{1}{2}} \right\| = \sup_{i\in N} \left(\frac{\lambda_i}{\lambda_i + \lambda} \right)^{\frac{1}{2}} \le 1.$$

For the second term in (12), using the concentration result between $T_{X\lambda}$ and T_λ, it can also be bounded by a constant. The main challenge of the proof is bounding the third term. We will show that the third term in (12) can be rewritten as

$$\begin{aligned} &\left\| T_\lambda^{-\frac{1}{2}} \left[(g_Z - (T_X + \lambda + T - T) f_\lambda) \right] \right\|_{\mathcal{H}} \\ &= \left\| T_\lambda^{-\frac{1}{2}} \left[(g_Z - T_X f_\lambda) - (T + \lambda) f_\lambda + T f_\lambda \right] \right\|_{\mathcal{H}} \\ &= \left\| T_\lambda^{-\frac{1}{2}} \left[(g_Z - T_X f_\lambda) - (g - T f_\lambda) \right] \right\|_{\mathcal{H}}. \end{aligned} \tag{13}$$

The form of (13) allows us to use the well-known Bernstein type inequality, which controls the difference between a sum of i.i.d. random variables and its expectation. Traditionally, this step requires the boundedness of f_ρ^*. We refine this procedure by a truncation method and prove that with high probability,

$$(13) \le \ln \frac{2}{\delta} C \frac{\lambda^{-\frac{1}{2\beta}}}{\sqrt{n}} = \ln \frac{2}{\delta} C n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}}.$$

Specifically, we consider two subset of $\mathcal{X}$: $\Omega_1 = \{x \in \Omega : |f_\rho^*(x)| \le t\}$ and $\Omega_2 = \mathcal{X} \backslash \Omega_1$ where t is allowed to diverge as $n \to \infty$. When choosing t as a proper order of n, we show that on the one hand, the norm in Ω_1 is still upper bounded by $n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}}$; on the other hand, the norms in Ω_2 vanishes with a probability tending to 1 as $n \to \infty$. We argue that such t exists by taking advantage of the L^q-integrability of f_ρ^*, which is required in the statement of Theorem 3.1.

4. Experiments

In our experiments, we aim to verify that for functions $f_\rho^* \in [\mathcal{H}]^s$ but not in L^∞, the KRR estimator can still achieve the convergence rate $n^{-\frac{s\beta}{s\beta+1}}$. We show the results for both Sobolev RKHS, and general RKHS with uniformly bounded eigenfunctions mentioned in Remark 3.3.

4.1. Experiments in Sobolev space

Suppose that $\mathcal{X} = [0, 1]$ and the marginal distribution μ is the uniform distribution on $[0, 1]$. We consider the RKHS $\mathcal{H} = H^1(\mathcal{X})$ to be the Sobolev space with smoothness 1. Section 3.1 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

$$f^*(x) = \sum_{k=1}^{\infty} \frac{1}{k^{s+0.5}} \left(\sin(2k\pi x) + \cos(2k\pi x) \right), \tag{14}$$

for some $0 < s < \frac{1}{\beta} = 0.5$. We will show in Appendix F that the series in (14) 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., $k(x, y) = \frac{1}{\sinh 1} \cosh(1 - \max(x, y)) \cosh(1 - \min(x, y))$.

We consider the following data generation procedure:

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

where f^* is numerically approximated by the first 1000 terms in (14) with $s = 0.4$, $x \sim \mathcal{U}[0, 1]$ and $\epsilon \sim \mathcal{N}(0, 1)$. We choose the regularization parameter as $\lambda = cn^{-\frac{\beta}{s\beta+1}} = cn^{-\frac{10}{9}}$ 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 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.

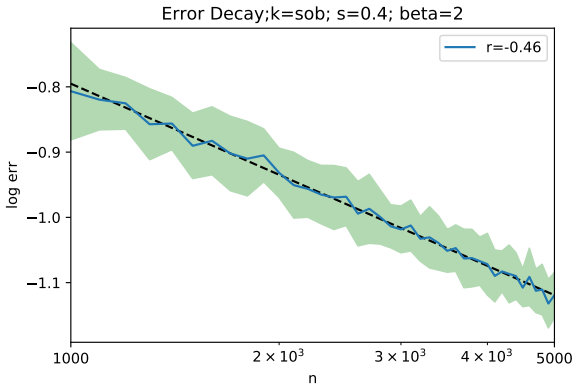


Figure 1. Error decay curve of Sobolev RKHS. Both axes are logarithmic. The curve shows the average generalization errors over 50 trials; the region within one standard deviation is shown in green. The dashed black line is computed using logarithmic least-squares, and the slope represents the convergence rate r.

We try different values of c, Figure 1 presents the convergence curve under the best choice of c. It can be concluded that the convergence rate of KRR's generalization error is 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 also did another experiment that used cross validation to choose the regularization parameter. Figure 3 in Appendix F shows a similar result as Figure 1. We refer to Appendix F for more details of the experiments.

Remark 4.1. This setting is similar to Pillaud-Vivien et al. (2018). The difference is that we choose the source condition s to be smaller than α_0 so that $f^* \notin L^\infty(\mathcal{X})$.

4.2. Experiments in general RKHS

Suppose that $\mathcal{X} = [0, 1]$ and the marginal distribution μ is the uniform distribution on $[0, 1]$. It is well known that the following RKHS

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

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 $\beta = 2$, the eigenfunctions are uniformly bounded, and the embedding index is $\alpha_0 = \frac{1}{\beta}$ (see Remark 3.3). We construct a function in $[\mathcal{H}]^s \backslash L^\infty$ by

$$f^*(x) = \sum_{k=1}^{\infty} \frac{1}{k^{s+0.5}} e_{2k-1}(x), \tag{15}$$

for some $0 < s < \frac{1}{\beta} = 0.5$. We will show in Appendix F that the series in (15) converges on $(0, 1)$. Since $e_{2k-1}(1) \equiv 1$, we also have $f^* \notin L^\infty(\mathcal{X})$.

We use the same data generation procedure as Section 4.1:

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

where f^* is numerically approximated by the first 1000 terms in (15) with $s = 0.4$, $x \sim \mathcal{U}[0, 1]$ and $\epsilon \sim \mathcal{N}(0, 1)$.

Figure 2 presents the convergence curve under the best choice of c. It can also be concluded that the convergence rate of KRR's generalization error is indeed approximately equal to $n^{-\frac{s\beta}{s\beta+1}} = n^{-\frac{4}{9}}$. Figure 4 in Appendix F shows the result of using cross validation.

5. Conclusion and Discussion

This paper considers the convergence rate and the minimax optimality of kernel ridge regression, especially in the misspecified case. When the true function f_ρ^* falls into a less-smooth ($s \leq \alpha_0$) interpolation space of the RKHS

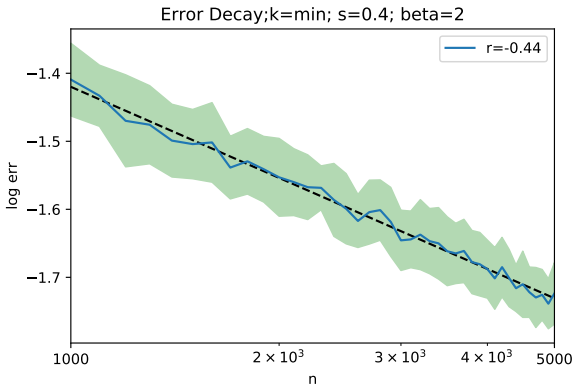


Figure 2. Error decay curve of general RKHS. Both axes are logarithmic. The curve shows the average generalization errors over 50 trials; the region within one standard deviation is shown in green. The dashed black line is computed using logarithmic least-squares, and the slope represents the convergence rate r.

$[\mathcal{H}]^s \nsubseteq L^\infty$, it has been an outstanding problem for years that whether the boundedness assumption of f_ρ^* in the proof can be removed or weakened so that KRR is optimal for all $0 < s \leq 2$. This paper proves that for Sobolev RKHS, the boundedness assumption can be directly removed. This result implies that, on the one hand, the desired upper bound of the convergence rate holds for all the functions in $[\mathcal{H}]^s$ (which can only be proved for $[\mathcal{H}]^s \cap L^\infty$ before); on the other hand, KRR is minimax optimal for all source conditions $0 < s \leq 2$ (which can only be proved for $\frac{1}{\beta} < s \leq 2$ before). Together with the saturation effect of KRR (Li et al., 2023) when $s > 2$, all convergence behaviors of Sobolev RKHS are understood. For general RKHS, we also prove that the boundedness assumption can be replaced by a L^q-integrability assumption which turns out to be reasonable, at least for RKHS with uniformly bounded eigenfunctions. We verify the results through experiments for both Sobolev RKHS and general RKHS.

If an RKHS $\mathcal{H}$ has embedding index $\alpha_0 = 1$, considering the embedding index will just recover the results of (Lin et al., 2018). Note that we assume $\alpha_0 \in (0, 1)$ throughout this paper, on which we will perform refined analysis. Technical tools in this paper can be extended to general spectral regularization algorithms and $[\mathcal{H}]^\gamma$-generalization error. Another direct open question is to discuss whether the L^q-integrability assumption in Theorem 3.1 holds for general RKHS, which we conjecture to be true.

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 magnitudes. At present, their results all rely on a Gaussian design assumption (or some variation), which is a very strong assumption. Nevertheless, their empirical results are enlightening and attract interest in the study of the learning curves of kernel ridge regression. 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. KRR is also connected with Gaussian process regression (Kanagawa et al., 2018). Jin et al. (2021) claimed to establish the learning curves for Gaussian process regression and thus for KRR. However, there is a gap in their proof where an essential covering argument is missing: their Corollary 20 provides a high probability bound for fixed ν, while the proof of their Lemma 40 and Lemma 41 mistakenly use Corollary 20 to assert that the bound holds simultaneously for all ν.

Since Jacot et al. (2018) introduced the NTK , kernel regression has become a natural surrogate for the neural networks in the 'lazy trained regime'. Our work is also motivated by our empirical studies in the neural networks and NTK regressions. Specifically, we found that the source condition with respect to the neural tangent kernel (or other frequently used kernels) is relatively small ($s \ll 1$) for some frequently used real datasets (MNIST, CIFAR-10). This observation urged us to determine the optimal convergence rate when the RKHS is misspecified.

Acknowledgements

This research was partially supported by Beijing Natural Science Foundation (Grant Z190001), the National Natural Science Foundation of China (Grant 11971257), National Key R&D Program of China (2020AAA0105200), and Beijing Academy of Artificial Intelligence.

References

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

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

Bauer, F., Pereverzyev, S., and Rosasco, L. On regularization algorithms in learning theory. *Journal of complexity*, 23(1):52–72, 2007.

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

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

Caponnetto, A. Optimal rates for regularization operators in learning theory. 2006.

Caponnetto, A. and de Vito, E. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7:331–368, 2007.

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

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

Cui, H., Loureiro, B., Krzakala, F., and Zdeborov'a, L. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. In *NeurIPS*, 2021.

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

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

Fischer, S.-R. and Steinwart, I. Sobolev norm learning rates for regularized least-squares algorithms. *Journal of Machine Learning Research*, 21:205:1–205:38, 2020.

Gerfo, L. L., Rosasco, L., Odone, F., Vito, E. D., and Verri, A. Spectral algorithms for supervised learning. *Neural Computation*, 20(7):1873–1897, 2008.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Jin, H., Banerjee, P. K., and Montúfar, G. Learning curves for gaussian process regression with power-law priors and targets. *ArXiv*, abs/2110.12231, 2021.

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

Kanagawa, M., Hennig, P., Sejdinovic, D., and Sriperumbudur, B. K. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*, 2018.

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

Li, Y., Zhang, H., and Lin, Q. On the saturation effect of kernel ridge regression. In *The Eleventh International Conference on Learning Representations*, 2023.

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

Lin, J. and Cevher, V. Optimal convergence for distributed learning with stochastic gradient methods and spectral algorithms. *Journal of Machine Learning Research*, 21: 147–1, 2020.

Lin, J., Rudi, A., Rosasco, L., and Cevher, V. Optimal rates for spectral algorithms with least-squares regression over Hilbert spaces. *Applied and Computational Harmonic Analysis*, 48:868–890, 2018.

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

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

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

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

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

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

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

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

Steinwart, I. and Christmann, A. Support vector machines. In *Information Science and Statistics*, 2008.

Steinwart, I. and Scovel, C. Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs. *Constructive Approximation*, 35(3):363–417, 2012.

Steinwart, I., Hush, D., and Scovel, C. Optimal rates for regularized least squares regression. In *COLT*, pp. 79–93, 2009.

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

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

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

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

A. Proof of Theorem 3.1

Throughout the proof, we denote

$$T_\lambda = T + \lambda; \;\; T_{X\lambda} = T_X + \lambda, \tag{16}$$

where λ is the regularization parameter and T, T_X is defined by (3), (7). 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.

In addition, denote the effective dimension

$$\mathcal{N}(\lambda) = \mathrm{tr}\big(T(T+\lambda)^{-1}\big) = \sum_{i\in N} \frac{\lambda_i}{\lambda_i + \lambda}, \tag{17}$$

since the EDR of $\mathcal{H}$ is β, Lemma E.5 shows that $\mathcal{N}(\lambda) \asymp \lambda^{-\frac{1}{\beta}}$.

The following lemma will be frequently used, so we present it at the beginning of our proof.

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

$$\sup_{t\geq 0} \frac{t^s}{t+\lambda} \leq \lambda^{s-1}. \tag{18}$$

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

$$\left(\frac{t}{\lambda}\right)^s \leq \frac{t}{\lambda} + 1 = \frac{t+\lambda}{\lambda}. \tag{19}$$

□

A.1. Proof of the approximation error

The following theorem gives the bound of $[\mathcal{H}]^\gamma$-norm of $f_\lambda - f_\rho^*$ when $0 \leq \gamma \leq s$. As a special case, the approximation error $\left\|f_\lambda - f_\rho^*\right\|_{L^2}$ follows from the result when $\gamma = 0$.

Theorem A.2. *Suppose that Assumption 2.3 holds for $0 < s \leq 2$. Denote $f_\lambda = (T+\lambda)^{-1} g$, then for any $\lambda > 0$ and $0 \leq \gamma \leq s$, we have*

$$\left\|f_\lambda - f_\rho^*\right\|_{[\mathcal{H}]^\gamma} \leq R\lambda^{\frac{s-\gamma}{2}}. \tag{20}$$

Proof. Suppose that $f_\rho^* = \sum_{i\in N} a_i e_i$. Recall that $f_\lambda = (T+\lambda)^{-1} g = (T+\lambda)^{-1} T f_\rho^*$, we have

$$\begin{aligned}
\left\|f_\lambda - f_\rho^*\right\|_{[\mathcal{H}]^\gamma}^2 &= \left\|\sum_{i\in N} \frac{\lambda}{\lambda+\lambda_i} a_i e_i(\cdot)\right\|_{[\mathcal{H}]^\gamma}^2 = \sum_{i\in N} \left(\frac{\lambda}{\lambda+\lambda_i}\right)^2 \lambda_i^{-\gamma} a_i^2 \\
&= \sum_{i\in N} \left(\frac{\lambda \lambda_i^{\frac{s-\gamma}{2}}}{\lambda+\lambda_i}\right)^2 \lambda_i^{-s} a_i^2 \\
&\leq \left(\lambda \sup_{i\in N} \frac{\lambda_i^{\frac{s-\gamma}{2}}}{\lambda+\lambda_i}\right)^2 \sum_{i\in N} \lambda_i^{-s} a_i^2 \\
&\leq \lambda^{\frac{s-\gamma}{2}} \|f_\rho^*\|_{[\mathcal{H}]^s},
\end{aligned} \tag{21}$$

where we use Lemma A.1 for the last inequality and the $[\mathcal{H}]^s$ norm defined by (5). Then the theorem follows from Assumption 2.3, i.e., $\|f_\rho^*\|_{[\mathcal{H}]^s} \leq R$.

□

A.2. Proof of the estimation error

Theorem A.3. *Suppose that Assumption 2.1, 2.2, 2.3 and 2.4 hold for* $0 < s \leq 2$ *and* $\frac{1}{\beta} \leq \alpha_0 < 1$. *Suppose that* $f_\rho^* \in L^q(\mathcal{X}, \mu)$ *and* $\|f_\rho^*\|_{L^q(\mathcal{X},\mu)} \leq C_q < \infty$ *for some* $q > \frac{2(s\beta+1)}{2+(s-\alpha_0)\beta}$. *Then in the case of* $s + \frac{1}{\beta} > \alpha_0$, *by choosing* $\lambda \asymp n^{-\frac{\beta}{s\beta+1}}$, *for any fixed* $\delta \in (0,1)$, *when* n *is sufficiently large, with probability at least* $1-\delta$, *we have*

$$\left\|\hat{f}_\lambda - f_\lambda\right\|_{L^2} \leq \ln\frac{4}{\delta} C n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}}. \tag{22}$$

where C *is a constant that only depends on* $\kappa, R, L, \sigma, C_q$.

Proof. First, rewrite the estimator error as follows

$$\left\|\hat{f}_\lambda - f_\lambda\right\|_{L^2} = \left\|T^{\frac{1}{2}}\left(\hat{f}_\lambda - f_\lambda\right)\right\|_{\mathcal{H}} \leq \left\|T^{\frac{1}{2}}T_\lambda^{-\frac{1}{2}}\right\| \cdot \left\|T_\lambda^{\frac{1}{2}}T_{X\lambda}^{-1}T_\lambda^{\frac{1}{2}}\right\| \cdot \left\|T_\lambda^{-\frac{1}{2}}\left(g_Z - T_{X\lambda}f_\lambda\right)\right\|_{\mathcal{H}}. \tag{23}$$

For the first term in (23), we have

$$\left\|T^{\frac{1}{2}}T_\lambda^{-\frac{1}{2}}\right\| = \sup_{i\in N}\left(\frac{\lambda_i}{\lambda_i+\lambda}\right)^{\frac{1}{2}} \leq 1. \tag{24}$$

For the second term in (23), for any $\alpha > \alpha_0$, using Proposition D.5 and we known that when λ, n satisfy that

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

we have

$$a := \|T_\lambda^{-\frac{1}{2}}(T - T_X)T_\lambda^{-\frac{1}{2}}\| \leq \frac{4}{3}u + \sqrt{2u} \leq \frac{2}{3}. \tag{26}$$

with probability as least $1 - \frac{\delta}{2}$. Therefore,

$$\begin{aligned}
\left\|T_\lambda^{\frac{1}{2}}T_{X\lambda}^{-1}T_\lambda^{\frac{1}{2}}\right\| &= \left\|\left(T_\lambda^{-\frac{1}{2}}(T_X+\lambda)T_\lambda^{-\frac{1}{2}}\right)^{-1}\right\| \\
&= \left\|\left(I - T_\lambda^{-\frac{1}{2}}(T_X - T)T_\lambda^{-\frac{1}{2}}\right)^{-1}\right\| \\
&\leq \sum_{k=0}^{\infty}\left\|T_\lambda^{-\frac{1}{2}}(T_X - T)T_\lambda^{-\frac{1}{2}}\right\|^k \\
&\leq \sum_{k=0}^{\infty}\left(\frac{2}{3}\right)^k \leq 3,
\end{aligned} \tag{27}$$

with probability as least $1 - \frac{\delta}{2}$, where I is the identity operator $\mathcal{H} \to \mathcal{H}$. Note that since we have $s + \frac{1}{\beta} > \alpha_0$, there always exists $\alpha_0 < \alpha < s + \frac{1}{\beta}$ such that (25) is satisfied when $\lambda \asymp n^{-\frac{\beta}{s\beta+1}}$ and n is sufficiently large .

For the third term in (23), it can be rewritten as

$$\begin{aligned}
\left\|T_\lambda^{-\frac{1}{2}}\left(g_Z - T_{X\lambda}f_\lambda\right)\right\|_{\mathcal{H}} &= \left\|T_\lambda^{-\frac{1}{2}}\left[\left(g_Z - \left(T_X + \lambda + T - T\right)f_\lambda\right)\right]\right\|_{\mathcal{H}} \\
&= \left\|T_\lambda^{-\frac{1}{2}}\left[\left(g_Z - T_X f_\lambda\right) - \left(T+\lambda\right)f_\lambda + Tf_\lambda\right]\right\|_{\mathcal{H}} \\
&= \left\|T_\lambda^{-\frac{1}{2}}\left[\left(g_Z - T_X f_\lambda\right) - \left(g - Tf_\lambda\right)\right]\right\|_{\mathcal{H}}.
\end{aligned} \tag{28}$$

Using Proposition D.4,with probability at least $1 - \frac{\delta}{2}$, we have

$$\left\|T_\lambda^{-\frac{1}{2}}\left[\left(g_Z - T_X f_\lambda\right) - \left(g - Tf_\lambda\right)\right]\right\|_{\mathcal{H}} \leq \ln\frac{4}{\delta}C\frac{\lambda^{-\frac{1}{2\beta}}}{\sqrt{n}} = \ln\frac{4}{\delta}Cn^{-\frac{1}{2}\frac{s\beta}{s\beta+1}}. \tag{29}$$

Plugging (24), (27) and (29) into (23) and we finish the proof. □

A.3. Proof of Theorem 3.1

Using the approximation-estimation error decomposition,

$$\left\|\hat{f}_\lambda - f_\rho^*\right\|_{L^2} \leq \left\|\hat{f}_\lambda - f_\lambda\right\|_{L^2} + \left\|f_\lambda - f_\rho^*\right\|_{L^2}, \tag{30}$$

together with Theorem A.2 and Theorem A.3 and we finish the proof of Theorem 3.1.

B. Proof of Theorem 3.6

Note that the RKHS $\mathcal{H}$ is defined as the (fractional) Sobolev space $H^r(\mathcal{X})$, which is regardless of the marginal distribution μ. But the definition of interpolation space (4) is dependent on μ. When μ has Lebesgue density $0 < c \leq p(x) \leq C$, Fischer & Steinwart (2020, (14)) shows that

$$L^2(\mathcal{X}, \mu) \cong L^2(\mathcal{X}, \nu), \tag{31}$$

and

$$[H^r(\mathcal{X})]_\mu^s \cong \left(L_2(\mathcal{X}, \mu), [H^r(\mathcal{X})]_\mu^1\right)_{s,2} \cong \left(L_2(\mathcal{X}, \nu), [H^r(\mathcal{X})]_\nu^1\right)_{s,2} \cong [H^r(\mathcal{X})]_\nu^s \cong H^{rs}(\mathcal{X}), \tag{32}$$

where we denote $[H^r(\mathcal{X})]_\mu^s$ as the interpolation space of $H^r(\mathcal{X})$ under marginal distribution μ. So we can also apply Proposition 3.5 to $[H^r(\mathcal{X})]_\mu^s$ and the embedding property of it is the same as $[H^r(\mathcal{X})]^s$. Denote $\beta = \frac{2r}{d}$, since $\alpha_0 = \frac{1}{\beta}$, for any $0 < s \leq 2$, we have

$$\text{Assumption 2.2; } s + \frac{1}{\beta} > \alpha_0; \; q > \frac{2(s\beta + 1)}{2 + (s - \alpha_0)\beta} = 2 \tag{33}$$

hold simultaneously.

Therefore, all the assumptions in Theorem 3.1 are satisfied, and the proof follows from applying Theorem 3.1.

C. Proof of Proposition 2.9

We will construct a family of probability distributions on $\mathcal{X} \times \mathcal{Y}$ and apply Lemma E.6. Recall that μ is a probability distribution on $\mathcal{X}$ such that Assumption 2.1 is satisfied. Denote the class of functions

$$B^s(R) = \left\{f \in [\mathcal{H}]^s : \|f\|_{[\mathcal{H}]^s} \leq R\right\}, \tag{34}$$

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

$$y = f(x) + \epsilon, \;\; x \sim \mu, \tag{35}$$

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 Proposition 2.9. (Assumption 2.1 and 2.3 are satisfied obviously. Assumption 2.4 follows from results of moments of Gaussian random variables, see, e.g., Fischer & Steinwart (2020, Lemma 21)).

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

$$\sum_{k=1}^m \left|\omega_k^{(i)} - \omega_k^{(j)}\right| \geq \frac{m}{8}, \quad \forall 0 \leq i < j \leq M. \tag{36}$$

For $\epsilon = C_0 m^{-s\beta-1}$, define the functions $f_i, i = 1, 2, \cdots, M$ as

$$f_i := \epsilon^{1/2} \sum_{k=1}^m \omega_k^{(i)} e_{m+k}. \tag{37}$$

Since

$$\|f_i\|_{[\mathcal{H}]^s} = \epsilon \sum_{k=1}^m \lambda_{m+k}^{-s} \left(\omega_k^{(i)}\right)^2 \leq \epsilon \sum_{k=1}^m \lambda_{2m}^{-s} \leq 2^{s\beta} c\epsilon \sum_{k=1}^m m^{s\beta} \leq 2^{s\beta} c\epsilon m^{s\beta+1} = 2^{s\beta} c C_0, \tag{38}$$

Where c in (38) represents the constant in Assumption 2.1. So if C_0 is small such that

$$2^{s\beta} c C_0 \leq R, \tag{39}$$

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

Using Lemma E.7, we have

$$\begin{aligned} \mathrm{KL}\left(\rho_{f_i}^n, \rho_{f_0}^n\right) &= \frac{n}{2\bar{\sigma}^2} \|f_i\|_{L^2(\mathcal{X},\mu)}^2 \\ &= \frac{n\epsilon}{2\bar{\sigma}^2} \sum_{k=1}^{m} \left(\omega_k^{(i)}\right)^2 \\ &\leq \frac{n\epsilon m}{2\bar{\sigma}^2} = \frac{n}{2\bar{\sigma}^2} C_0 m^{-s\beta}. \end{aligned} \tag{40}$$

Recall that $M \geq 2^{m/8}$ implies $\ln M \geq \frac{\ln 2}{8} m$. For a fixed $a \in (0, \frac{1}{8})$, since $m = n^{\frac{1}{s\beta+1}}$, letting

$$\mathrm{KL}\left(\rho_{f_i}^n, \rho_{f_0}^n\right) \leq \frac{n}{2\bar{\sigma}^2} C_0 m^{-s\beta} \leq a \frac{\ln 2}{8} m \leq a \ln M, \tag{41}$$

we have

$$C_0 \leq \frac{\bar{\sigma}^2 \ln 2}{4} a. \tag{42}$$

So we can choose $C_0 = c' a$ such that (39) and (42) are satisfied.

Denote $\left\{\rho_{f_i}^n, f_i \in B^s(R)\right\}$ as a family of probability distribution index by f_i, then (41) implies the second condition in Lemma E.6 holds. Further, using (36), we have

$$d\left(f_i, f_j\right)^2 = \|f_i - f_j\|_{L^2}^2 = \epsilon \sum_{k=1}^{m} \left(\omega_k^{(i)} - \omega_k^{(j)}\right)^2 \geq \frac{\epsilon m}{8} = \frac{c' a}{8} m^{-s\beta} \geq c' a n^{-\frac{s\beta}{s\beta+1}}, \tag{43}$$

where c' is a constant independent of n.

Applying Lemma E.6 to (41) and (43), we have

$$\inf_{\hat{f}_n} \sup_{f \in B^s(R)} \mathbb{P}_{\rho_f} \left\{ \left\| \hat{f}_n - f \right\|_{L^2}^2 \geq c' a n^{-\frac{s\beta}{s\beta+1}} \right\} \geq \frac{\sqrt{M}}{1+\sqrt{M}} \left(1 - 2a - \sqrt{\frac{2a}{\ln M}}\right). \tag{44}$$

When n is sufficiently large so that M is sufficiently large, the probability in the R.H.S. of (44) 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 (44) shows that there exists a constant C, 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 $1 - \delta$,

$$\left\| \hat{f} - f \right\|_{L^2}^2 \geq C \delta n^{-\frac{s\beta}{s\beta+1}}. \tag{45}$$

So we finish the proof.

D. Useful propositions for upper bounds

This proposition bounds the L^∞ norm of f_λ when $s \leq \alpha_0$.

Proposition D.1. *Suppose that Assumption 2.1, 2.2 and 2.3 hold for $0 < s \leq \alpha_0$ and $\frac{1}{\beta} \leq \alpha_0 < 1$. Denote $f_\lambda = (T+\lambda)^{-1} g$, then for any $\lambda > 0$ and any $\alpha > \alpha_0$, we have*

$$\|f_\lambda\|_{L^\infty} \leq M_\alpha \|f_\rho^*\|_{[\mathcal{H}]^s} \lambda^{-\frac{\alpha - s}{2}}. \tag{46}$$

Proof. Suppose that $f_\rho^* = \sum_{i\in N} a_i e_i$. Since $s \le \alpha_0$ and $\alpha > \alpha_0$, we have

$$\begin{aligned}\|f_\lambda\|_{[\mathcal{H}]^\alpha}^2 &= \sum_{i\in N}\left(\frac{\lambda_i^{1-\frac{\alpha}{2}}}{\lambda+\lambda_i}\right)^2 a_i^2 \\ &= \sum_{i\in N}\left(\frac{\lambda_i^{1-\frac{\alpha-s}{2}}}{\lambda+\lambda_i}\right)^2 \lambda_i^{-s} a_i^2 \\ &\le \|f_\rho^*\|_{[\mathcal{H}]^s}^2 \lambda^{s-\alpha},\end{aligned} \tag{47}$$

where we use Lemma A.1 for the last inequality. Further, using $\|[\mathcal{H}]^\alpha \hookrightarrow L^\infty(\mathcal{X},\mu)\| = M_\alpha$ by Assumption 2.2, we have $\|f_\lambda\|_{L^\infty} \le M_\alpha \|f_\lambda\|_{[\mathcal{H}]^\alpha} \le M_\alpha \|f_\rho^*\|_{[\mathcal{H}]^s}\lambda^{-\frac{\alpha-s}{2}}$.

□

The following proposition is an application of the classical Bernstein type inequality but considering a truncation version of f_ρ^*, which will bring refined analysis compared to previous work.

Proposition D.2. *Suppose that Assumption 2.1, 2.2, 2.3 and 2.4 hold for $0 < s \le 2$ and $\frac{1}{\beta} \le \alpha_0 < 1$. Denote $\xi_i = \xi(x_i, y_i) = T_\lambda^{-\frac{1}{2}}(K_{x_i}y_i - T_{x_i}f_\lambda)$ and $\Omega_0 = \{x \in \Omega : |f_\rho^*(x)| \le t\}$. Then for any $\alpha > \alpha_0$ and all $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\left\|\frac{1}{n}\sum_{i=1}^n \xi_i I_{x_i\in\Omega_0} - \mathbb{E}\xi_x I_{x\in\Omega_0}\right\|_{\mathcal{H}} \le \ln\frac{2}{\delta}\left(\frac{C_1\lambda^{-\frac{\alpha}{2}}}{n}\cdot\tilde{M} + \frac{C_2\mathcal{N}^{\frac{1}{2}}(\lambda)}{\sqrt{n}} + \frac{C_3\lambda^{-\frac{\alpha-s}{2}}}{\sqrt{n}}\right), \tag{48}$$

where $\tilde{M} = M_\alpha R\lambda^{-\frac{\alpha-s}{2}} + t + L$, and L is the constant in (2.4). $C_1 = 8\sqrt{2}M_\alpha, C_2 = 8\sigma, C_3 = 8\sqrt{2}M_\alpha R$.

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 E.4, we need to bound the m-th moment of $\xi(x,y)I_{x\in\Omega_0}$.

$$\begin{aligned}\mathbb{E}\,\|\xi(x,y)I_{x\in\Omega_0}\|_{\mathcal{H}}^m &= \mathbb{E}\left\|T_\lambda^{-\frac{1}{2}}K_x(y-f_\lambda(x))I_{x\in\Omega_0}\right\|_{\mathcal{H}}^m \\ &\le \mathbb{E}\Big(\left\|T_\lambda^{-\frac{1}{2}}k(x,\cdot)\right\|_{\mathcal{H}}^m \mathbb{E}\big(|(y-f_\lambda(x))I_{x\in\Omega_0}|^m \mid x\big)\Big).\end{aligned} \tag{49}$$

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

$$\begin{aligned}|y-f_\lambda(x)|^m &\le 2^{m-1}\left(\left|f_\lambda(x)-f_\rho^*(x)\right|^m + \left|f_\rho^*(x)-y\right|^m\right) \\ &= 2^{m-1}\left(\left|f_\lambda(x)-f_\rho^*(x)\right|^m + |\epsilon|^m\right).\end{aligned} \tag{50}$$

Plugging (50) into (49), we have

$$\mathbb{E}\,\|\xi(x,y)I_{x\in\Omega_0}\|_{\mathcal{H}}^m \;\le\; 2^{m-1}\mathbb{E}\Big(\left\|T_\lambda^{-\frac{1}{2}}k(x,\cdot)\right\|_{\mathcal{H}}^m \left|(f_\lambda(x)-f_\rho^*(x))I_{x\in\Omega_0}\right|^m\Big) \tag{51}$$

$$+\, 2^{m-1}\mathbb{E}\Big(\left\|T_\lambda^{-\frac{1}{2}}k(x,\cdot)\right\|_{\mathcal{H}}^m \mathbb{E}\big(|\epsilon\, I_{x\in\Omega_0}|^m \mid x\big)\Big) \tag{52}$$

Now we begin to bound (52). Note that we have proved in Lemma E.1 that for μ-almost $x \in \mathcal{X}$,

$$\left\|T_\lambda^{-\frac{1}{2}}k(x,\cdot)\right\|_{\mathcal{H}} \le M_\alpha\lambda^{-\frac{\alpha}{2}}; \tag{53}$$

In addition, we also have

$$\begin{aligned}\mathbb{E}\left\|T_\lambda^{-\frac{1}{2}}k(x,\cdot)\right\|_{\mathcal{H}}^2 &= \mathbb{E}\Big\|\sum_{i\in N}(\frac{1}{\lambda_i+\lambda})^{\frac{1}{2}}\lambda_i e_i(x)e_i(\cdot)\Big\|_{\mathcal{H}}^2\\ &= \mathbb{E}\Big(\sum_{i\in N}\frac{\lambda_i}{\lambda_i+\lambda}e_i^2(x)\Big)\\ &= \sum_{i\in N}\frac{\lambda_i}{\lambda_i+\lambda}\\ &= \mathcal{N}(\lambda).\end{aligned} \tag{54}$$

So we have

$$\mathbb{E}\left\|T_\lambda^{-\frac{1}{2}}k(x,\cdot)\right\|_{\mathcal{H}}^m \le \sup_{x\in\mathcal{X}}\left\|T_\lambda^{-\frac{1}{2}}k(x,\cdot)\right\|_{\mathcal{H}}^{m-2}\cdot\mathbb{E}\left\|T_\lambda^{-\frac{1}{2}}k(x,\cdot)\right\|_{\mathcal{H}}^2 \le \big(M_\alpha\lambda^{-\frac{\alpha}{2}}\big)^{m-2}\mathcal{N}(\lambda). \tag{55}$$

Using Assumption 2.4, we have

$$\mathbb{E}\left(|\epsilon I_{x\in\Omega_0}|^m \mid x\right) \le \mathbb{E}\left(|\epsilon|^m \mid x\right) \le \frac{1}{2}m!\sigma^2L^{m-2}, \quad \mu\text{-a.e. } x\in\mathcal{X}, \tag{56}$$

so we get the upper bound of (52), i.e.,

$$(52) \le \frac{1}{2}m!\left(\sqrt{2}\sigma\mathcal{N}^{\frac{1}{2}}(\lambda)\right)^2(2M_\alpha\lambda^{-\frac{\alpha}{2}}L)^{m-2}. \tag{57}$$

Now we begin to bound (51).

(1) When $s\le\alpha_0$, using the definition of Ω_0 and Proposition D.1, we have

$$\|(f_\lambda-f_\rho^*)I_{x\in\Omega_0}\|_{L^\infty} \le \|f_\lambda\|_{L^\infty}+\|f_\rho^*I_{x\in\Omega_0}\|_{L^\infty} \le M_\alpha R\lambda^{-\frac{\alpha-s}{2}}+t := M. \tag{58}$$

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

$$\|(f_\lambda-f_\rho^*)I_{x\in\Omega_0}\|_{L^\infty} \le M_\alpha\|f_\lambda-f_\rho^*\|_{[\mathcal{H}]^\alpha} \le M_\alpha R\lambda^{-\frac{\alpha-s}{2}} < M. \tag{59}$$

Therefore, for all $0<s\le 2$ we have

$$\|(f_\lambda-f_\rho^*)I_{x\in\Omega_0}\|_{L^\infty} \le M. \tag{60}$$

In addition, using Theorem A.2 for $\gamma=0$, we also have

$$\mathbb{E}|(f_\lambda(x)-f_\rho^*(x))I_{x\in\Omega_0}|^2 \le \mathbb{E}|f_\lambda(x)-f_\rho^*(x)|^2 \le (R\lambda^{\frac{s}{2}})^2. \tag{61}$$

So we get the upper bound of (51), i.e.,

$$\begin{aligned}(51) &\le 2^{m-1}(M_\alpha\lambda^{-\frac{\alpha}{2}})^m\cdot\|(f_\lambda-f_\rho^*)I_{x\in\Omega_0}\|_{L^\infty}^{m-2}\cdot\mathbb{E}|(f_\lambda(x)-f_\rho^*(x))I_{x\in\Omega_0}|^2\\ &\le 2^{m-1}(M_\alpha\lambda^{-\frac{\alpha}{2}})^m\cdot M^{m-2}\cdot(R\lambda^{\frac{s}{2}})^2\\ &\le \frac{1}{2}m!\big(2M_\alpha\lambda^{-\frac{\alpha}{2}}M\big)^{m-2}\big(2M_\alpha R\lambda^{-\frac{\alpha-s}{2}}\big)^2.\end{aligned} \tag{62}$$

Denote

$$\begin{aligned}\tilde{L} &= 2M_\alpha(M+L)\lambda^{-\frac{\alpha}{2}}\\ \tilde{\sigma} &= 2M_\alpha R\lambda^{-\frac{\alpha-s}{2}}+\sqrt{2}\sigma\mathcal{N}^{\frac{1}{2}}(\lambda),\end{aligned} \tag{63}$$

then we have $\mathbb{E}\,\|\xi(x,y)I_{x\in\Omega_0}\|_{\mathcal{H}}^m \le \frac{1}{2}m!\tilde{\sigma}^2\tilde{L}^{m-2}$. Using Lemma E.4, we finish the proof. □

Remark D.3. In fact, when we later applying Proposition D.2 in the proof of Proposition D.4, the truncation method in this proposition 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 proposition.

Based on Proposition D.2, the following Proposition will give an upper bound of $\left\| T_\lambda^{-\frac{1}{2}} \left[(g_Z - T_X f_\lambda) - (g - T f_\lambda) \right] \right\|_{\mathcal{H}}$.

Proposition D.4. *Suppose that Assumption 2.1, 2.2, 2.3 and 2.4 hold for $0 < s \leq 2$ and $\frac{1}{\beta} \leq \alpha_0 < 1$. Suppose that $f_\rho^* \in L^q(\mathcal{X}, \mu)$ and $\|f_\rho^*\|_{L^q(\mathcal{X},\mu)} \leq C_q < \infty$ for some $q > \frac{2(s\beta+1)}{2+(s-\alpha_0)\beta}$. Then in the case of $s + \frac{1}{\beta} > \alpha_0$, by choosing $\lambda \asymp n^{-\frac{\beta}{s\beta+1}}$, for any fixed $\delta \in (0,1)$, when n is sufficiently large, with probability at least $1-\delta$, we have*

$$\left\| T_\lambda^{-\frac{1}{2}} \left[(g_Z - T_X f_\lambda) - (g - T f_\lambda) \right] \right\|_{\mathcal{H}} \leq \ln \frac{2}{\delta} C \frac{\lambda^{-\frac{1}{2\beta}}}{\sqrt{n}} = \ln \frac{2}{\delta} C n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}}, \tag{64}$$

where C is a constant that only depends on $\kappa, R, L, \sigma, C_q$.

Proof. Denote $\xi_i = \xi(x_i, y_i) = T_\lambda^{-\frac{1}{2}}(K_{x_i} y_i - T_{x_i} f_\lambda)$ and $\xi_x = \xi(x, y) = T_\lambda^{-\frac{1}{2}}(K_x y - T_x f_\lambda)$, then (64) is equivalent to

$$\left\| \frac{1}{n} \sum_{i=1}^n \xi_i - \mathbb{E}\xi_x \right\|_{\mathcal{H}} \leq \ln \frac{2}{\delta} C \frac{\lambda^{-\frac{1}{2\beta}}}{\sqrt{n}} = \ln \frac{2}{\delta} C n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}}. \tag{65}$$

Consider the subset $\Omega_1 = \{x \in \Omega : |f_\rho^*(x)| \leq t\}$ and $\Omega_2 = \mathcal{X} \backslash \Omega_1$. Since $\|f_\rho^*\|_{L^q(\mathcal{X},\mu)} \leq C_q$, we have

$$P(x \in \Omega_2) = P\Big(|f_\rho^*(x)| > t \Big) \leq \frac{\mathbb{E}|f_\rho^*(x)|^q}{t^q} \leq \frac{(C_q)^q}{t^q}. \tag{66}$$

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

$$\left\| \frac{1}{n} \sum_{i=1}^n \xi_i - \mathbb{E}\xi_x \right\|_{\mathcal{H}} \leq \left\| \frac{1}{n} \sum_{i=1}^n \xi_i I_{x_i \in \Omega_1} - \mathbb{E}\xi_x I_{x \in \Omega_1} \right\|_{\mathcal{H}} + \left\| \frac{1}{n} \sum_{i=1}^n \xi_i I_{x_i \in \Omega_2} \right\|_{\mathcal{H}} + \left\| \mathbb{E}\xi_x I_{x \in \Omega_2} \right\|_{\mathcal{H}}. \tag{67}$$

Given $s + \frac{1}{\beta} > \alpha_0$, here we firstly fixed an α such that

$$\alpha_0 < \alpha < s + \frac{1}{\beta}. \tag{68}$$

For the first term in (67), denoted as I, using Theorem D.2, for all $\delta \in (0,1)$, with probability at least $1-\delta$, we have

$$\mathrm{I} \leq \ln \frac{2}{\delta} \left(\frac{C_1 \lambda^{-\frac{\alpha}{2}}}{n} \cdot \tilde{M} + \frac{C_2 \mathcal{N}^{\frac{1}{2}}(\lambda)}{\sqrt{n}} + \frac{C_3 \lambda^{-\frac{\alpha-s}{2}}}{\sqrt{n}} \right), \tag{69}$$

where $\tilde{M} = M_\alpha R \lambda^{-\frac{\alpha-s}{2}} + t + L$. Recalling that $s + \frac{1}{\beta} > \alpha_0$, simple calculation shows that by choosing $\lambda = n^{-\frac{\beta}{s\beta+1}}$,

- the second term in (69):

$$\frac{C_2 \mathcal{N}^{\frac{1}{2}}(\lambda)}{\sqrt{n}} \asymp \frac{\lambda^{-\frac{1}{2\beta}}}{\sqrt{n}} = n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}}; \tag{70}$$

- the third term in (69):

$$\frac{C_3 \lambda^{-\frac{\alpha-s}{2}}}{\sqrt{n}} \asymp n^{\frac{1}{2}(\frac{\alpha}{s+1/\beta}-1)} \cdot n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}} \lesssim n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}}; \tag{71}$$

- the first term in (69):

$$\frac{C_1 \lambda^{-\frac{\alpha}{2}}}{n} \cdot \tilde{M} \asymp \frac{\lambda^{-\frac{\alpha}{2}}}{n} \lambda^{-\frac{\alpha-s}{2}} + \frac{\lambda^{-\frac{\alpha}{2}}}{n} \cdot t + \frac{\lambda^{-\frac{\alpha}{2}}}{n} \cdot L. \tag{72}$$

Further calculations show that

$$\frac{\lambda^{-\frac{\alpha}{2}}}{n}\lambda^{-\frac{\alpha-s}{2}} = n^{\frac{\alpha}{s+1/\beta}-1}\cdot n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}} \lesssim n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}}, \tag{73}$$

and

$$\frac{\lambda^{-\frac{\alpha}{2}}}{n} = n^{\frac{1}{2}\frac{\alpha\beta-s\beta-2}{s\beta+1}}\cdot n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}} \lesssim n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}}. \tag{74}$$

To make (72) $\lesssim n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}}$ when $\lambda = n^{-\frac{\beta}{s\beta+1}}$, letting $\frac{\lambda^{-\frac{\alpha}{2}}}{n}\cdot t \le n^{-\frac{1}{2}\frac{s\beta}{s\beta+1}}$, we have the first restriction of t:

$$\textbf{(R1)}: \quad t \le n^{\frac{1}{2}(1+\frac{1-\alpha\beta}{s\beta+1})}. \tag{75}$$

That is to say, if we choose $t \le n^{\frac{1}{2}(1+\frac{1-\alpha\beta}{s\beta+1})}$, we have

$$\mathrm{I} \le \ln\frac{2}{\delta}C\frac{\lambda^{-\frac{1}{2\beta}}}{\sqrt{n}} = \ln\frac{2}{\delta}Cn^{-\frac{1}{2}\frac{s\beta}{s\beta+1}}. \tag{76}$$

For the second term in (67), denoted as II, we have

$$\begin{aligned}
\tau_n := P(\mathrm{II} > \frac{\lambda^{-\frac{1}{2\beta}}}{\sqrt{n}}) &\le P\Big(\exists x_i \text{ s.t. } x_i \in \Omega_2\Big) = 1 - P\Big(x_i \notin \Omega_2, \forall i = 1,2,\cdots,n\Big) \\
&= 1 - P\Big(x \notin \Omega_2\Big)^n \\
&= 1 - P\Big(|f_\rho^*(x)| \le t\Big)^n \\
&\le 1 - \Big(1 - \frac{(C_q)^q}{t^q}\Big)^n.
\end{aligned} \tag{77}$$

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

$$\textbf{(R2)}: \quad t \gg n^{\frac{1}{q}}, \text{ or } n^{\frac{1}{q}} = o(t). \tag{78}$$

For the third term in (67), denoted as III. Since we have already known that $\|T_\lambda^{-\frac{1}{2}}k(x,\cdot)\|_{\mathcal{H}} \le \lambda^{-\frac{\alpha}{2}}$, μ-a.e. $x \in \mathcal{X}$, so

$$\begin{aligned}
\mathrm{III} &\le \mathbb{E}\|\xi_x I_{x\in\Omega_2}\|_{\mathcal{H}} \le \mathbb{E}\Big[\|T_\lambda^{-\frac{1}{2}}k(x,\cdot)\|_{\mathcal{H}}\cdot\big|\big(y - f_\lambda(x)\big)I_{x\in\Omega_2}\big|\Big] \\
&\le \lambda^{-\frac{\alpha}{2}}\mathbb{E}\big|\big(y - f_\lambda(x)\big)I_{x\in\Omega_2}\big| \\
&\le \lambda^{-\frac{\alpha}{2}}\Big(\mathbb{E}\big|\big(f_\rho^*(x) - f_\lambda(x)\big)I_{x\in\Omega_2}\big| + \mathbb{E}\big|\big(f_\rho^*(x) - y\big)I_{x\in\Omega_2}\big|\Big) \\
&\le \lambda^{-\frac{\alpha}{2}}\Big(\mathbb{E}\big|\big(f_\rho^*(x) - f_\lambda(x)\big)I_{x\in\Omega_2}\big| + \mathbb{E}\big|\epsilon\cdot I_{x\in\Omega_2}\big|\Big).
\end{aligned} \tag{79}$$

Using Cauchy-Schwarz and the bound of approximation error (Theorem A.2), we have

$$\mathbb{E}\big|\big(f_\rho^*(x) - f_\lambda(x)\big)I_{x\in\Omega_2}\big| \le \Big(\big\|f_\rho^* - f_\lambda\big\|_{L^2}\Big)^{\frac{1}{2}}\cdot(P(x\in\Omega_2))^{\frac{1}{2}} \le R\lambda^{\frac{s}{2}}C_q^{\frac{q}{2}}t^{-\frac{q}{2}}. \tag{80}$$

In addition, we have

$$\mathbb{E}\big|\epsilon\cdot I_{x\in\Omega_2}\big| = \mathbb{E}\left(\mathbb{E}\big|\epsilon\cdot I_{x\in\Omega_2}\big|\,\Big|\,x\right) \le \sigma\mathbb{E}\,|I_{x\in\Omega_2}| \le \sigma(C_q)^q t^{-q}. \tag{81}$$

Plugging (80) and (81) into (79), we have

$$\mathrm{III} \le RC_q^{\frac{q}{2}}\lambda^{-\frac{\alpha-s}{2}}t^{-\frac{q}{2}} + \sigma(C_q)^q\lambda^{-\frac{\alpha}{2}}t^{-q}. \tag{82}$$

Comparing (82) with $C_3\frac{\lambda^{-\frac{\alpha-s}{2}}}{\sqrt{n}}$ and $C_1\frac{\lambda^{-\frac{\alpha}{2}}}{n}$ in (69). We know that if $t \ge n^{\frac{1}{q}}$, (79) $\le C\frac{\lambda^{-\frac{1}{2\beta}}}{\sqrt{n}} = Cn^{-\frac{1}{2}\frac{s\beta}{s\beta+1}}$. So the third term will not give further restriction of t.

To sum up, if we choose t such that restrictions (75) and (78) are satisfied, then we can prove that (65) 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 (65) is satisfied with probability at least $1-\delta$.

Note that, such t exists if

$$\frac{1}{q} < \frac{1}{2}(1+\frac{1-\alpha\beta}{s\beta+1}) \Longleftrightarrow q > \frac{2(s\beta+1)}{2+(s-\alpha)\beta}. \tag{83}$$

Recalling that for (68), we only assume there exists α satisfying $\alpha_0 < \alpha < s+\frac{1}{\beta}$, so such t exists if and only if

$$\frac{1}{q} < \frac{1}{2}(1+\frac{1-\alpha_0\beta}{s\beta+1}) \Longleftrightarrow q > \frac{2(s\beta+1)}{2+(s-\alpha_0)\beta}. \tag{84}$$

which is what we assume in the theorem. □

Proposition D.5. *Suppose that the embedding index is α_0. Then for any $\alpha > \alpha_0$ and all $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\|T_\lambda^{-\frac{1}{2}}(T-T_X)T_\lambda^{-\frac{1}{2}}\| \le \frac{4M_\alpha^2\lambda^{-\alpha}}{3n}B + \sqrt{\frac{2M_\alpha^2\lambda^{-\alpha}}{n}B}, \tag{85}$$

where

$$B = \ln\frac{4\mathcal{N}(\lambda)(\|T\|+\lambda)}{\delta\|T\|}. \tag{86}$$

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

$$\|A_i\| = \|T_\lambda^{-\frac{1}{2}}TT_\lambda^{-\frac{1}{2}}\| + \|T_\lambda^{-\frac{1}{2}}T_{x_i}T_\lambda^{-\frac{1}{2}}\| \le 2M_\alpha^2\lambda^{-\alpha}. \tag{87}$$

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

$$\mathbb{E}A_i^2 \preceq \mathbb{E}\left[T_\lambda^{-\frac{1}{2}}T_{x_i}T_\lambda^{-\frac{1}{2}}\right]^2. \tag{88}$$

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

$$\mathbb{E}A_i^2 \preceq \mathbb{E}\left[T_\lambda^{-\frac{1}{2}}T_{x_i}T_\lambda^{-\frac{1}{2}}\right]^2 \preceq \mathbb{E}\left[M_\alpha^2\lambda^{-\alpha}\cdot T_\lambda^{-\frac{1}{2}}T_{x_i}T_\lambda^{-\frac{1}{2}}\right] = M_\alpha^2\lambda^{-\alpha}T_\lambda^{-1}T, \tag{89}$$

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

$$\begin{aligned}
\|V\| &= M_\alpha^2\lambda^{-\alpha}\frac{\lambda_1}{\lambda_1+\lambda} = M_\alpha^2\lambda^{-\alpha}\frac{\|T\|}{\|T\|+\lambda} \le M_\alpha^2\lambda^{-\alpha};\\
\mathrm{tr}V &= M_\alpha^2\lambda^{-\alpha}\mathcal{N}(\lambda);\\
\frac{\mathrm{tr}V}{\|V\|} &= \frac{\mathcal{N}(\lambda)(\|T\|+\lambda)}{\|T\|}.
\end{aligned} \tag{90}$$

Using Lemma E.3 to A_i, V, we finish the proof. □

E. Auxiliary lemma

E.1. Lemmas for upper bound

The following lemma is where we take advantage of the embedding index and embedding property in Assumption 2.2.

Lemma E.1. *Suppose that the embedding index is α_0. Then for any $\alpha > \alpha_0$, for μ-almost $x \in \mathcal{X}$, we have*

$$\|T_\lambda^{-\frac{1}{2}}k(x,\cdot)\|_{\mathcal{H}}^2 \le M_\alpha^2\lambda^{-\alpha}. \tag{91}$$

Proof. Recalling that $\|[\mathcal{H}]^\alpha \hookrightarrow L^\infty(\mathcal{X})\| = M_\alpha$, we have

$$\begin{aligned}\|T_\lambda^{-\frac{1}{2}}k(x,\cdot)\|_{\mathcal{H}}^2 &= \Big\|\sum_{i\in N}(\frac{1}{\lambda_i+\lambda})^{\frac{1}{2}}\lambda_i e_i(x)e_i(\cdot)\Big\|_{\mathcal{H}}^2\\ &= \sum_{i\in N}\frac{\lambda_i}{\lambda_i+\lambda}e_i^2(x)\\ &= \big[\sum_{i\in N}\lambda_i^\alpha e_i^2(x)\big]\sup_{i\in N}\frac{\lambda_i^{1-\alpha}}{\lambda_i+\lambda}\\ &\le M_\alpha^2\lambda^{-\alpha},\end{aligned} \tag{92}$$

where we use Lemma A.1 for the last inequality, and we finish the proof. □

Lemma E.1 has a direct corollary.

Corollary E.2. *Suppose that the embedding index is α_0. Then for any $\alpha > \alpha_0$, for μ-almost $x \in \mathcal{X}$, we have*

$$\|T_\lambda^{-\frac{1}{2}}T_xT_\lambda^{-\frac{1}{2}}\| \le M_\alpha^2\lambda^{-\alpha}. \tag{93}$$

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

$$\begin{aligned}T_\lambda^{-\frac{1}{2}}T_xT_\lambda^{-\frac{1}{2}}f &= T_\lambda^{-\frac{1}{2}}K_xK_x^*T_\lambda^{-\frac{1}{2}}f\\ &= T_\lambda^{-\frac{1}{2}}K_x\langle k(x,\cdot),T_\lambda^{-\frac{1}{2}}f\rangle_{\mathcal{H}}\\ &= T_\lambda^{-\frac{1}{2}}K_x\langle T_\lambda^{-\frac{1}{2}}k(x,\cdot),f\rangle_{\mathcal{H}}\\ &= \langle T_\lambda^{-\frac{1}{2}}k(x,\cdot),f\rangle_{\mathcal{H}}\cdot T_\lambda^{-\frac{1}{2}}k(x,\cdot).\end{aligned} \tag{94}$$

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

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

Lemma E.3. *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 $\|A_i\| \le L$ 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*

$$\left\|\frac{1}{n}\sum_{i=1}^n A_i\right\| \le \frac{2L\beta}{3n} + \sqrt{\frac{2\|V\|\beta}{n}}, \quad \beta = \ln\frac{4\mathrm{trV}}{\delta\|V\|}. \tag{95}$$

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

Lemma E.4 (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*

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

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

$$\left\|\frac{1}{n}\sum_{i=1}^n\xi_i - \mathbb{E}\xi\right\|_H \le 4\sqrt{2}\ln\frac{2}{\delta}\left(\frac{L}{n}+\frac{\sigma}{\sqrt{n}}\right). \tag{96}$$

Lemma E.5. *If $\lambda_i \asymp i^{-\beta}$, we have*

$$\mathcal{N}(\lambda) \asymp \lambda^{-\frac{1}{\beta}}. \tag{97}$$

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

$$\mathcal{N}(\lambda) = \sum_{i=1}^{\infty} \frac{\lambda_i}{\lambda_i + \lambda} \leq \sum_{i=1}^{\infty} \frac{C i^{-\beta}}{C i^{-\beta} + \lambda} = \sum_{i=1}^{\infty} \frac{C}{C + \lambda i^{\beta}} \tag{98}$$

$$\leq \int_0^{\infty} \frac{C}{\lambda x^{\beta} + C} \mathrm{d}x = \lambda^{-\frac{1}{\beta}} \int_0^{\infty} \frac{C}{y^{\beta} + C} \mathrm{d}y \leq C_1 \lambda^{-\frac{1}{\beta}}. \tag{99}$$

for some constant C_1. Similarly, we have

$$\mathcal{N}(\lambda) \geq C_2 \lambda^{-\frac{1}{\beta}}, \tag{100}$$

for some constant C_2. □

E.2. Lemmas for minimax 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 E.6. *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*

$$\frac{1}{M} \sum_{j=1}^{M} K(P_j, P_0) \leq a \log M, \tag{101}$$

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

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} P_\theta(d(\hat{\theta}, \theta) \geq s) \geq \frac{\sqrt{M}}{1 + \sqrt{M}} \left(1 - 2a - \sqrt{\frac{2a}{\log M}}\right). \tag{102}$$

Lemma E.7. *Suppose that μ is a distribution on $\mathcal{X}$ and $f_i \in L^2(\mathcal{X}, \mu)$. Suppose that*

$$y = f_i(x) + \epsilon, \quad i = 1, 2, \tag{103}$$

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

$$K(P_1, P_2) \coloneqq \int_{\Omega} \log\left(\frac{\mathrm{d}P_1}{\mathrm{d}P_2}\right) \mathrm{d}P_1, \tag{104}$$

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

$$\mathrm{KL}(\rho_1^n, \rho_2^n) = n\mathrm{KL}(\rho_1, \rho_2) = \frac{n}{2\sigma^2} \|f_1 - f_2\|^2_{L^2(\mathcal{X}, d\mu)}, \tag{105}$$

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

$$\mathrm{KL}\left(N\left(f_1(x), \sigma^2\right), N\left(f_2(x), \sigma^2\right)\right) = \frac{1}{2\sigma^2} |f_1(x) - f_2(x)|^2. \tag{106}$$

□

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

Lemma E.8. *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* $\omega^{(0)} = (0, \cdots, 0)$,

$$d_{Ham}\left(\omega^{(i)}, \omega^{(j)}\right) := \sum_{k=1}^{m} \left|\omega_k^{(i)} - \omega_k^{(j)}\right| \geq \frac{m}{8}, \quad \forall 0 \leq i < j \leq M, \tag{107}$$

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

F. Details of experiments

F.1. Experiments in Sobolev RKHS

First, we prove that the series in (14) 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

$$\begin{aligned}
&-2\sin(\pi x)\left(\sin(2\pi x) + \sin(4\pi x) + \cdots + \sin(2N\pi x)\right) \\
&= \left[\cos(2\pi + \pi)x - \cos(2\pi - \pi)x\right] + \left[\cos(4\pi + \pi)x - \cos(4\pi - \pi)x\right] \\
&\quad + \cdots + \left[\cos(2N\pi + \pi)x - \cos(2N\pi - \pi)x\right] \\
&= \cos(2N\pi + \pi)x - \cos \pi x.
\end{aligned} \tag{108}$$

So we have

$$\left|\left(\sin(2\pi x) + \sin(4\pi x) + \cdots + \sin(2N\pi x)\right)\right| = \frac{\left|\cos(2N\pi + \pi)x - \cos \pi x\right|}{\left|2\sin(\pi x)\right|}; \tag{109}$$

Similarly, we have

$$\left|\left(\cos(2\pi x) + \cos(4\pi x) + \cdots + \cos(2N\pi x)\right)\right| = \frac{\left|\sin(2N\pi + \pi)x - \sin \pi x\right|}{\left|2\sin(\pi x)\right|}. \tag{110}$$

Note that (109) and (110) 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 (14) 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 $(0,1)$.

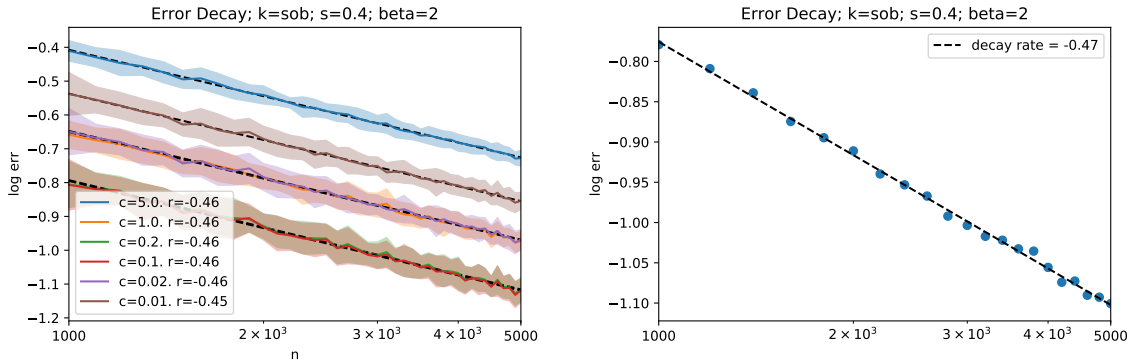
In Figure 3 (a), we present the results of different choices of c for $\lambda = cn^{-\frac{\beta}{s\beta+1}}$ in the experiment of Section 4.2. Figure 2 corresponds to the curve $c = 0.1$ in Figure 3 (a), which has the smallest generalization error. In Figure 3 (b), we use 5-fold cross validation to choose the regularization parameter in KRR and present the logarithmic errors and sample sizes. Again, we use logarithmic least-squares to compute the convergence rate r, which is still approximately equal to $n^{-\frac{s\beta}{s\beta+1}} = n^{-\frac{4}{9}}$.

F.2. Experiments in general RKHS

First, we prove that the series in (15) 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 $e_{2k-1}(x)$,

$$\begin{aligned}
&-2\sin(\pi x)\left(\sin\left(\frac{\pi x}{2}\right) + \sin\left(\frac{5\pi x}{2}\right) + \cdots + \sin\left(\frac{(4N-3)\pi x}{2}\right)\right) \\
&= \left[\cos\left(\pi + \frac{\pi}{2}\right)x - \cos\left(\pi - \frac{\pi}{2}\right)x\right] + \left[\cos\left(\frac{5\pi}{2} + \pi\right)x - \cos\left(\frac{5\pi}{2} - \pi\right)x\right] \\
&\quad + \cdots + \left[\cos\left(\frac{(4N-3)\pi}{2} + \pi\right)x - \cos\left(\frac{(4N-3)\pi}{2} - \pi\right)x\right] \\
&= \cos\left(\frac{(4N-1)\pi}{2}\right)x - \cos\frac{\pi}{2}x.
\end{aligned} \tag{111}$$



(a) fixed c (b) cross validation

Figure 3. Error decay curves of Sobolev RKHS. Both axes are logarithmic. (a) The curves show the average generalization errors of different c over 50 trials; and the regions within one standard deviation are shown in the corresponding colors. (b) The scatters show the average generalization errors obtained by 5-fold cross validation over 50 trials. In both (a) and (b), the dashed black lines are computed using logarithmic least-squares, and the slopes represent the convergence rates r.

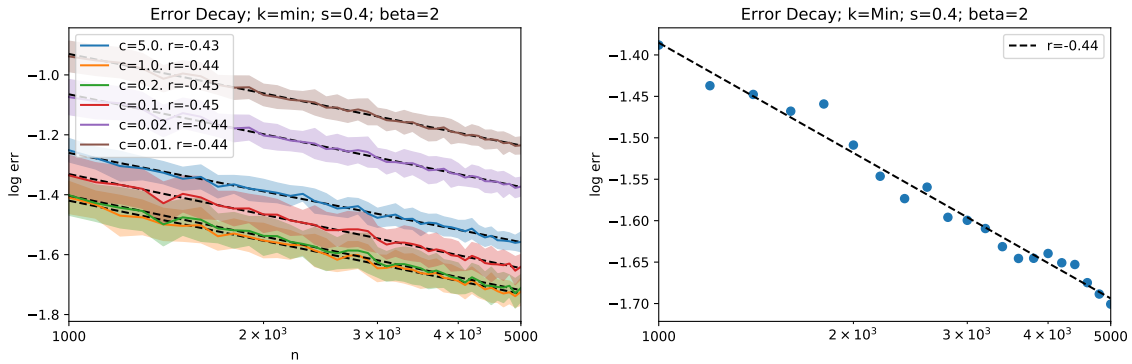
So we have

$$\left|\sin\left(\frac{\pi x}{2}\right)+\sin\left(\frac{5\pi x}{2}\right)+\cdots+\sin\left(\frac{(4N-3)\,\pi x}{2}\right)\right| = \frac{\left|\cos\left(\frac{(4N-1)\pi}{2}\right)x-\cos\frac{\pi}{2}x\right|}{|2\sin(\pi x)|}, \tag{112}$$

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 (15) 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 4 (a), we present the results of different choices of c for $\lambda = cn^{-\frac{\beta}{s\beta+1}}$ in the experiment of Section 4.2. Figure 2 corresponds to the curve $c = 1.0$ in Figure 4 (a), which has the smallest generalization error. In Figure 4 (b), we use 5-fold cross validation to choose the regularization parameter in KRR and present the logarithmic errors and sample sizes. Again, we use logarithmic least-squares to compute the convergence rate r, which is still approximately equal to $n^{-\frac{s\beta}{s\beta+1}} = n^{-\frac{4}{9}}$.



(a) fixed c (b) cross validation

Figure 4. Error decay curves of general RKHS.