Journal of Machine Learning Research 25 (2024) 1-88 Submitted 5/23; Revised 7/24; Published 8/24

Random Smoothing Regularization in Kernel Gradient Descent Learning

Liang Ding * LIANG_DING@FUDAN.EDU.CN
Fudan University
Shanghai, China

Tianyang Hu HUTIANYANG.UP@OUTLOOK.COM
Huawei Noah's Ark Lab
Shenzhen, China

Jiahang Jiang JJIANGBC@CONNECT.UST.HK
The Hong Kong University of Science and Technology
Hong Kong SAR, China

Donghao Li DLIBF@CONNECT.UST.HK
The Hong Kong University of Science and Technology
Hong Kong SAR, China

Wenjia Wang WENJIAWANG@HKUST-GZ.EDU.CN
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
The Hong Kong University of Science and Technology
Hong Kong SAR, China

Yuan Yao YUANY@UST.HK
The Hong Kong University of Science and Technology
Hong Kong SAR, China

Editor: Jean-Philippe Vert

Abstract

Random smoothing data augmentation is a unique form of regularization that can prevent overfitting by introducing noise to the input data, encouraging the model to learn more generalized features. Despite its success in various applications, there has been a lack of systematic study on the regularization ability of random smoothing. In this paper, we aim to bridge this gap by presenting a framework for random smoothing regularization that can adaptively and effectively learn a wide range of ground truth functions belonging to the classical Sobolev spaces. Specifically, we investigate two underlying function spaces: the Sobolev space of low intrinsic dimension, which includes the Sobolev space in D-dimensional Euclidean space or low-dimensional sub-manifolds as special cases, and the mixed smooth Sobolev space with a tensor structure. By using random smoothing regular-

*. The authors' names are sorted alphabetically, and the corresponding author is Wenjia Wang.

ization as novel convolution-based smoothing kernels, we can attain optimal convergence rates in these cases using a kernel gradient descent algorithm, either with early stopping or weight decay. It is noteworthy that our estimator can adapt to the structural assumptions of the underlying data and avoid the curse of dimensionality. This is achieved through various choices of injected noise distributions such as Gaussian, Laplace, or general polynomial noises, allowing for broad adaptation to the aforementioned structural assumptions of the underlying data. The convergence rate depends only on the effective dimension, which may be significantly smaller than the actual data dimension. We conduct numerical experiments on simulated data to validate our theoretical results.

Keywords: random smoothing, regularization, kernel gradient descent, early stopping, weight decay

1. Introduction

Random smoothing data augmentation is a technique used to improve the generalization and robustness of machine learning models, particularly in the context of deep learning. This method involves adding random noise, such as Gaussian or Laplace noise, to the input data during the training process. The idea behind random smoothing is to make the model more robust to small perturbations in the input data, as the added noise simulates variations that may occur naturally in real-world data. This augmentation approach has proven to be an effective regularization technique, contributing to the empirical success of deep learning models across various applications. For instance, random flip, random crop, and color jitter can significantly improve the classification accuracy in natural images (Goodfellow et al., 2016; Shorten and Khoshgoftaar, 2019). Random smoothing has been proven effective for improving model robustness and generalization (Blum et al., 2020; Rosenfeld et al., 2020; Mehra et al., 2021; Wang et al., 2020; Gao et al., 2020). For example, random smoothing with Gaussian noise injection is introduced to address the adversarial vulnerability (Cohen et al., 2019; Salman et al., 2019), and by encouraging the feature map to be invariant under data augmentations, self-supervised contrastive learning methods (He et al., 2020; Chen et al., 2020; Grill et al., 2020; Chen and He, 2021; He et al., 2021) can achieve state-of-the-art performance for various downstream tasks.

Random smoothing can be viewed as a form of regularization (Grandvalet et al., 1997). Regularization techniques generally aim to reduce the complexity of a model, making it less prone to fitting the noise in the training data and, consequently, improving its performance on unseen data. Random smoothing can be considered an implicit form of regularization, as it does not directly modify the model's parameters or loss function, unlike explicit regularization techniques such as ℓ_1 or ℓ_2 regularization. Instead, it indirectly influences the model's behavior by altering the input data during training. By adding random noise to the input data, random smoothing forces the model to focus on the underlying structure of the data rather than memorizing specific instances. This leads to more robust and generalizable models that can better handle variations in real-world data. As a result, random smoothing acts as a regularizer, improving the model's ability to generalize from the training set to unseen data. Such a regularization perspective at least starts with Grandvalet et al. (1997). However, in spite of the empirical success of random smoothing in various applications, there is a lack of systematic research on the regularization effect of random smoothing in the literature.

In this paper, we address this gap by examining the classic nonparametric regression problem from the perspective of random smoothing regularization. In nonparametric regression, the primary objective is to uncover the functional relationship between input and output variables. By making appropriate assumptions about the underlying truth function and selecting the appropriate estimator, we focus on understanding the efficiency of the estimation, specifically, the rate at which the estimation error converges to zero as the sample size n increases. The optimal convergence rate is typically dictated by the problem's inherent complexity. The actual achievable convergence rates depend on the specific estimation methods employed. Among various techniques, we consider kernel methods that have been extensively investigated in the research literature (Wahba, 1990; Hastie et al., 2001).

In this study, we present a unified framework that can learn a wide range of D-dimensional ground truth functions belonging to the classical Sobolev spaces ($\mathcal{W}^{m_f}$) in an effective and adaptive manner. The framework incorporates random smoothing as a central component. Our hypothesis space is a reproducing kernel Hilbert space that is associated with a kernel function of smoothness denoted by m_0. Random smoothing regularization leads to a novel convolution between the kernel function and a probability density function for the injected input noise. This injected noise is governed by either short or long-tail distributions, namely Gaussian and polynomial (including Laplace) noises, respectively. The resulting convolution-based random smoothing kernel enables us to adapt to the smoothness of the target functions more efficiently. Notably, we establish that for any m_0 and m_f greater than $D/2$, optimal convergence rates can be achieved by utilizing random smoothing regularization and appropriate early stopping and/or weight decay techniques.

To be specific, we investigate two possible function spaces that may contain the target function. In Section 4.2, we analyze the Sobolev space with a low intrinsic dimension, which is denoted by d. This space covers both D-dimensional Euclidean spaces (when $d = D$) and low-dimensional sub-manifolds as specific examples. In Section 4.3, we explore the mixed smooth Sobolev spaces, which possess a tensor structure. Our principal findings are summarized below.

- In case of Sobolev space of low intrinsic dimensionality $d \leq D$:

  When using Gaussian random smoothing, an upper bound of the convergence rate is achieved at $n^{-m_f/(2m_f+d)}(\log n)^{D+1}$, which recovers the results presented in Hamm and Steinwart (2021a) and is hypothetically optimal up to a logarithmic factor. However, in contrast to Hamm and Steinwart (2021a), we present a different approach that allows us to analyze polynomial smoothing;

  When using polynomial random smoothing with data size adaptive smoothing degree, a convergence rate of $n^{-m_f/(2m_f+d)}(\log n)^{2m_f+1}$ is achieved, which is again, hypothetically optimal up to a logarithmic factor.

- In case of mixed smooth Sobolev spaces, using polynomial random smoothing of degree m_ε, a fast convergence rate of $n^{-2m_f/(2m_f+1)}(\log n)^{\frac{2m_f}{2m_f+1}\left(D-1+\frac{1}{2(m_0+m_\varepsilon)}\right)}$ is achieved, which is optimal up to a logarithmic factor.

To the best of our knowledge, such results have not been studied in the literature so far. They have various implications below.

First of all, these results enhance the convergence rates in the context of kernel ridge regression by incorporating random smoothing data augmentation with two other popular techniques, early stopping and weight decay. In kernel ridge regression, it is crucial to balance the smoothness of the kernel function (m_0) with that of the ground truth (m_f). In practice, it is common for m_0 to be unequal to m_f. In cases of mismatch, regularization becomes essential. Specifically, if $m_0 \in [m_f/2, \infty)$, the optimal convergence rate $n^{-m_f/(2m_f+D)}$ can be achieved by employing an appropriate ridge penalty strength. This result can be generalized to low intrinsic dimensionality $d \leq D$, where the hypothetically optimal convergence rate is $n^{-m_f/(2m_f+d)}$ (Hamm and Steinwart, 2021a). However, when the chosen kernel has a smoothness m_0 less than $m_f/2$, the optimal adaptation is not well studied in kernel ridge regression. To the best of our knowledge, only upper bounds are available in this case, and is not optimal. For example, Wang and Jing (2022) derived the upper error bound of the form $n^{-2m_0/(4m_0+d)}$ for $m_0 < m_f/2$. In comparison, similar convergence rate of the form $n^{-m_0/(2m_0+d)}$ can be achieved by distributed learning but without disjoint subset (Guo et al., 2017; Lin et al., 2017). In contrast, our findings demonstrate optimal adaptation for arbitrary m_0 and $m_f \geq D/2$ without such a constraint. This highlights the broad adaptation ability of random smoothing regularization.

Moreover, the optimal adaptation of polynomial random smoothing has an implication for neural networks via the (generalized) Laplace random smoothing. It is known that the training of neural networks, with enough overparametrization, can be characterized by kernel methods with a special family of kernels called the "neural tangent kernel" (NTK). Due to the low smoothness of the ReLU activation function, the corresponding NTK also has a low smoothness that is the same as a Laplace kernel (Chen and Xu, 2020; Geifman et al., 2020). To the best of our knowledge, the estimation error is at the rate $n^{-\frac{D}{2D-1}}$ (Hu et al., 2021). Our results, using the polynomial random smoothing with (generalized) Laplace distributions, show that the convergence rate can be improved, which sheds light on understanding non-smooth augmentations such as random crop and mask. Based on this understanding, numerical experiments with neural networks are conducted on simulated data to corroborate our theoretical results.

Finally, it is worth mentioning that with random smoothing, the convergence rates mentioned above can be obtained by early stopping. However, if one applies weight decay, the number of iterations can be reduced from polynomial(n) to polynomial($\log n$). Additionally, our estimator can adapt to the low-dimensional assumptions mentioned earlier, as the convergence rates depend on D at most logarithmically, alleviating the curse of dimensionality. It is also important to note that we do not employ the spectrum of integral operator technique (Yao et al., 2007; Lin et al., 2016; Lin and Rosasco, 2017), but instead use Fourier analysis, which provides a universal basis for kernels of different smoothness, and avoids imposing conditions on the eigenvalues and eigenfunctions of the kernel function. This is because there is no clear relationship between the low intrinsic dimension and the eigenvalues of the integral operator. Furthermore, our theoretical analysis can be applied to the widely used Matérn kernel functions.

The remainder of this paper is structured as follows. In Section 2, we provide a review of related works. Section 3 introduces the settings considered in this work, which include early stopping with a random smoothing kernel, as well as the conditions and assumptions utilized in this work. The main theoretical results are presented in Section 4, and numerical studies are conducted in Section 5. Conclusions and a discussion are provided in Section 7. Technical proofs are included in the Appendix.

2. Related Works

Various means of regularization have been proposed for kernel methods to better recover the underlying function, among which, ridge penalty and early stopping are the most popular. Kernel ridge regression has been extensively studied in the literature, see Blanchard and Mücke (2018); Dicker et al. (2017); Guo et al. (2017); Lin et al. (2017); Steinwart et al. (2009); Tuo et al. (2020); Wu et al. (2006) for example. Early stopping treats the number of training iterations as a hyperparameter in the optimization process, which has been extensively studied by the applied mathematics community (Dieuleveut and Bach, 2016; Yao et al., 2007; Pillaud-Vivien et al., 2018; Raskutti et al., 2014). Various forms of early stopping also have been studied including boosting (Zhang and Yu, 2005; Bartlett and Traskin, 2007), conjugate gradient algorithm (Blanchard and Krämer, 2016) and kernel gradient descent (Bühlmann and Yu, 2002; Caponnetto and Yao, 2010; Yao et al., 2007; Wei et al., 2017; Lin et al., 2016). Some works (e.g. Lin et al., 2016; Lin and Rosasco, 2017; Pillaud-Vivien et al., 2018) have explored early stopping by employing the integral operator induced by the kernel, imposing conditions on the eigenvalues and eigenfunctions of the kernel function. Smoothness or regularity of functions thus implicitly depends on the measure that defines the spectrum of the integral operator, whereas classical smoothness like Sobolev spaces is not explicitly handled.

In kernel regression with gradient descent, Raskutti et al. (2014) showed that early stopping and ridge penalty both can achieve the optimal convergence rate if the smoothness is well-specified. Yet, kernel ridge regression might suffer the "saturation issues" while early stopping does not (Engl et al., 1996; Yao et al., 2007). In regression problems, it is usually assumed that the domain of interest has a positive Lebesgue measure, while in practice, the data generating distribution is supported on some low-dimensional smooth sub-manifold (Scott and Nowak, 2006; Yang and Dunson, 2016; Ye and Zhou, 2008, 2009; Hamm and Steinwart, 2021b,a). Kernel methods can circumvent the curse of dimensionality and adapt to various low-dimensional assumptions of the underlying function. In particular, Hamm and Steinwart (2021b,a) generalized the manifold assumption by applying the box-counting dimension of the support of the data distribution, and derived upper bounds on the convergence rate of the prediction error. Another simplifying assumption is tensor product kernels (Gretton, 2015; Szabó and Sriperumbudur, 2017), whose product forms allow efficient computation of Gaussian process regression (Saatçi, 2012; Wilson and Nickisch, 2015; Ding and Zhang, 2022; Chen et al., 2022) and analysis of independent component (Bach and Jordan, 2002; Gretton et al., 2005, 2007). The RKHS induced by a tensor product kernel is simply tensored RKHS (Paulsen and Raghupathi, 2016, Theorem 5.11). Tensor product kernels we consider induce the tensored Sobolev spaces (Rieger and Wendland, 2017, Proposition 1).

For complicated high-dimensional data, deep learning models seem to perform extremely well, which has sparked numerous investigations into their generalization ability. As it turns out, the training of neural networks has deep connections to kernel methods with neural tangent kernels (NTK). Under proper initialization, training sufficiently wide DNN with gradient descent equates to kernel regression using NTK. First introduced by Jacot et al. (2018), the correspondence has been significantly extended (Du et al., 2018; Li and Liang, 2018; Arora et al., 2019a; Cao and Gu, 2020; Arora et al., 2019b; Li et al., 2019; Huang et al., 2020; Kanoh and Sugiyama, 2021; Hu et al., 2022). From the NTK point of view, ridge penalty and early stopping are also vital in training neural networks. The former is equivalent to weight decay (Hu et al., 2021), which is applied by default in training deep learning models for better generalization, so is early stopping (Prechelt, 1998). Zhang et al. (2021); Hardt et al. (2016) revealed that longer training can harm the generalization performance of deep models. Li et al. (2020); Bai et al. (2021) utilized early stopping to improve robustness to label noises.

Besides NTK, various data augmentation techniques in deep learning that are proven effective in improving model generalization can also provide inspiration for kernel methods. Grandvalet et al. (1997) studied from a regularization perspective how noise injection can improve generalization. Data augmentation is particularly important for handling natural images (Shorten and Khoshgoftaar, 2019), where horizontal flip, random crop, color jitter can significantly improve the classification accuracy. By applying the above augmentations, self-supervised contrastive learning methods (He et al., 2020; Chen et al., 2020; Grill et al., 2020; Chen and He, 2021; He et al., 2021) can achieve state-of-the-art performance for various downstream tasks. Randomized smoothing (Cohen et al., 2019; Salman et al., 2019) is a special data augmentation, first proposed to address the adversarial vulnerability (Goodfellow et al., 2014; Carlini and Wagner, 2017) of deep learning models. The key idea is to perturb the input with random noise injection and make predictions by aggregating the outputs from all augmented inputs. Random smoothing has been proven effective for improving model robustness and generalization (Rosenfeld et al., 2020; Mehra et al., 2021; Wang et al., 2020; Gao et al., 2020). Our proposed framework incorporates random smoothing, together with weight decay and early stopping, to provide a unified solution for the smoothness mismatch problem in kernel regression. It is worth clarifying the difference between our method and the "errors in variables" literature (Zhou et al., 2019; Wang et al., 2022; Cressie and Kornak, 2003; Cervone and Pillai, 2015). Though the formulations seem similar, i.e., the inputs in both cases are corrupted with noises, the two are fundamentally different. In our setting, both the input $\boldsymbol{x}$ and added noise $\boldsymbol{\varepsilon}$ are known (we control the noises in our estimator) while in the other setting, the input is noisy and only $\boldsymbol{x} + \boldsymbol{\varepsilon}$ is observed.

3. Random Smoothing Kernel Regression

In this section, we introduce the problem of interest, our methodology, and the necessary conditions used in this work.

3.1 Problem Setting

Suppose we have observed data $(\boldsymbol{x}_j, y_j)$ for $j = 1, ..., n$, which follows the relationship given by

$$y_j = f^*(\boldsymbol{x}_j) + \epsilon_j. \tag{1}$$

Here, $\boldsymbol{x}_j$'s are independent and identically distributed (i.i.d.) following a marginal distribution $P_{\mathbf{X}}$ with support $\mathrm{supp}(P_{\mathbf{X}}) = \Omega \subset \mathbb{R}^D$. The function $f^* \in \mathcal{H}(\Omega)$, where $\mathcal{H}(\Omega)$ denotes a function space, and ϵ_j's are i.i.d. noise variables with mean zero and finite variance. Our objective is to recover the function f^* based on the noisy observations.

In this work, we consider two cases. In the first case (Section 4.2), the function space $\mathcal{H}(\Omega)$ is a Sobolev space with smoothness m, denoted by $\mathcal{W}^m(\Omega)$, and the data is of low intrinsic dimension. In the second case (Section 4.3), the function space $\mathcal{H}(\Omega)$ is a tensor Sobolev space. Throughout this work, we assume without loss of generality that $P_{\mathbf{X}}$ follows a uniform distribution. Note that our theoretical analysis can be easily extended to the case where $P_{\mathbf{X}}$ is upper and lower bounded by positive constants. Specifically, suppose the density of $P_{\mathbf{X}}$, denoted by $p(\boldsymbol{x})$, satisfies $0 < c_1 \leq p(\boldsymbol{x}) \leq c_2 < \infty$, then it can be shown that $c_1\|f\|^2_{\mathrm{Unif}(\Omega)} \leq \|f\|^2_{P(\mathbf{X})} \leq c_2\|f\|^2_{\mathrm{Unif}(\Omega)}$, and our theoretical analysis can be mimicked. Furthermore, we can extend our results to unbounded regions by applying the truncation technique to light-tailed densities (e.g., sub-Gaussian densities).

In order to recover the function f^*, we use reproducing kernel Hilbert spaces (RKHSs). We briefly introduce the RKHSs and their relationship with Sobolev spaces in the following, and refer to Wendland (2004) and Adams and Fournier (2003) for details. Let $K : \Omega \times \Omega \to \mathbb{R}$ be a symmetric positive definite kernel function. Define the linear space

$$F_K(\Omega) = \left\{ \sum_{k=1}^{n} \beta_k K(\cdot, \boldsymbol{x}_k) : \beta_k \in \mathbb{R}, \boldsymbol{x}_k \in \Omega, n \in \mathbb{N} \right\}, \tag{2}$$

and equip this space with the bilinear form

$$\left\langle \sum_{k=1}^{n} \beta_k K(\cdot, \boldsymbol{x}_k), \sum_{j=1}^{m} \gamma_j K(\cdot, \boldsymbol{x}'_j) \right\rangle_K := \sum_{k=1}^{n} \sum_{j=1}^{m} \beta_k \gamma_j K(\boldsymbol{x}_k, \boldsymbol{x}'_j).$$

Then the reproducing kernel Hilbert space $\mathcal{H}_K(\Omega)$ generated by the kernel function K is defined as the closure of $F_K(\Omega)$ under the inner product $\langle \cdot, \cdot \rangle_K$, and the norm of $\mathcal{H}_K(\Omega)$ is $\|f\|_{\mathcal{H}_K(\Omega)} = \sqrt{\langle f, f \rangle_{\mathcal{H}_K(\Omega)}}$, where $\langle \cdot, \cdot \rangle_{\mathcal{H}_K(\Omega)}$ is induced by $\langle \cdot, \cdot \rangle_K$. The following theorem gives another characterization of the reproducing kernel Hilbert space when K is stationary, via the Fourier transform. Our notion of the Fourier transform is

$$\mathcal{F}(g)(\boldsymbol{\omega}) = (2\pi)^{-D/2} \int_{\mathbb{R}^D} g(\boldsymbol{x}) e^{-i\boldsymbol{\omega}^T \boldsymbol{x}} \mathrm{d}\boldsymbol{x},$$

for a function $g \in L_1(\mathbb{R}^D)$. Note that a kernel function K is said to be stationary if the value $K(\boldsymbol{x}, \boldsymbol{x}')$ only depends on the difference $\boldsymbol{x} - \boldsymbol{x}'$. Thus, we can write $K(\boldsymbol{x} - \boldsymbol{x}') := K(\boldsymbol{x}, \boldsymbol{x}')$. In this work, we only consider the stationary kernel due to the ease of mathematical treatment. Our theory can be generalized to the case where the kernel function is non-stationary but the corresponding RKHS is norm-equivalent to an RKHS generated by a stationary kernel. The general non-stationary kernel, albeit its flexibility, is out of the scope of this work, and will be pursued in the future.

Theorem 1 (Theorem 10.12 of Wendland, 2004) *Let K be a positive definite kernel function that is stationary, continuous, and integrable in $\mathbb{R}^D$. Define*

$$\mathcal{G} := \{f \in L_2(\mathbb{R}^D) \cap C(\mathbb{R}^D) : \mathcal{F}(f)/\sqrt{\mathcal{F}(K)} \in L_2(\mathbb{R}^D)\},$$

with the inner product

$$\langle f, g \rangle_{\mathcal{H}_K(\mathbb{R}^D)} = (2\pi)^{-d/2} \int_{\mathbb{R}^D} \frac{\mathcal{F}(f)(\boldsymbol{\omega})\overline{\mathcal{F}(g)(\boldsymbol{\omega})}}{\mathcal{F}(K)(\boldsymbol{\omega})} \mathrm{d}\boldsymbol{\omega}.$$

Then $\mathcal{G} = \mathcal{H}_K(\mathbb{R}^D)$, and both inner products coincide.

For $m > D/2$, the (fractional) Sobolev norm for function g on $\mathbb{R}^D$ is defined by

$$\|g\|^2_{\mathcal{W}^m(\mathbb{R}^D)} = \int_{\mathbb{R}^D} |\mathcal{F}(g)(\boldsymbol{\omega})|^2 (1 + \|\boldsymbol{\omega}\|_2^2)^m \mathrm{d}\boldsymbol{\omega}, \tag{3}$$

and the inner product of a Sobolev space $\mathcal{W}^m(\mathbb{R}^D)$ is defined by

$$\langle f, g \rangle_{\mathcal{W}^m(\mathbb{R}^D)} = \int_{\mathbb{R}^D} \mathcal{F}(f)(\boldsymbol{\omega})\overline{\mathcal{F}(g)(\boldsymbol{\omega})}(1 + \|\omega\|_2^2)^m \mathrm{d}\boldsymbol{\omega}.$$

Remark 2 *In this work, we are only interested in Sobolev spaces with $m > D/2$ because these spaces contain only continuous functions according to the Sobolev embedding theorem.*

Comparing Theorem 1 and (3), it can be seen that if

$$c_1(1 + \|\boldsymbol{\omega}\|_2^2)^{-m} \leq \mathcal{F}(K)(\boldsymbol{\omega}) \leq c_2(1 + \|\boldsymbol{\omega}\|_2^2)^{-m}, \forall \boldsymbol{\omega} \in \mathbb{R}^D,$$

for some two constants $c_1, c_2 > 0$, then $\mathcal{W}^m(\mathbb{R}^D)$ coincides with the reproducing kernel Hilbert space $\mathcal{H}_K(\mathbb{R}^D)$ with equivalent norms (also see Wendland, 2004, Corollary 10.13). By the extension theorem (DeVore and Sharpley, 1993), $\mathcal{H}_K(\Omega)$ also coincides with $\mathcal{W}^m(\Omega)$, and two norms are equivalent.

3.2 Random Smoothing Kernel Regression with Early Stopping

In this study, we systematically investigate the efficiency of random smoothing data augmentation, which is a widely used technique in deep learning, in improving the estimation efficiency (i.e., convergence rate) for $f^* \in \mathcal{H}(\Omega)$ without assuming any relationship between $\mathcal{H}(\Omega)$ and $\mathcal{H}_K(\Omega)$ and considering a wide context of Ω that may have Lebesgue measure zero. To overcome the lack of smoothness in $\mathcal{H}_K(\Omega)$, we construct N augmentations for each observed input point $\boldsymbol{x}_j$ by adding i.i.d. noise $\boldsymbol{\varepsilon}_{jk}$ with a continuous probability density function p_ε. We can generate $\boldsymbol{\varepsilon}_{jk}$ independently for each j, or we can generate $\boldsymbol{\varepsilon}_k$ for $k = 1, ..., N$, and apply them to all $\boldsymbol{x}_j$, $j = 1, ..., n$ simultaneously. While the latter is easier to implement, the former is easier to theoretically justify. Due to its lower computational complexity, we only consider the latter method in this work.

Remark 3 (Adding non-smooth noise and practical data augmentation techniques) *It should be noted that we do not assume p_ε to be Gaussian, and can be non-smooth. While applying Gaussian noise is a common practice, not all data augmentation techniques involve smooth noise, such as random crop, random mask, and random flip. In this work, we investigate various types of noise, including non-smooth Laplace noise and smooth Gaussian noise. Although adding non-smooth noise still cannot capture the effects of complex data augmentation techniques such as random mask or random crop, we aim to use it as a tool to gain insights into the success of these more complicated data augmentations.*

With augmented data, we proceed to the estimation of the function f^*. For any point $\boldsymbol{x} \in \Omega$, we obtain the estimator by computing the average of the function values evaluated at the N augmented inputs. Specifically, the estimator is constructed as

$$f(\boldsymbol{x}) = \frac{1}{N} \sum_{k=1}^{N} h(\boldsymbol{x} + \boldsymbol{\varepsilon}_k), \tag{4}$$

for $h \in \mathcal{H}_K(\Omega)$. By properties of the RKHS, f as in (4) is also inside $\mathcal{H}_K(\Omega)$. We consider the following l_2 loss function defined as

$$L_n(f) = \frac{1}{2n} \sum_{j=1}^{n} \left(f(\boldsymbol{x}_j) - y_j \right)^2, \tag{5}$$

or equivalently,

$$L_n(h) = \frac{1}{2n} \sum_{j=1}^{n} \left(\frac{1}{N} \sum_{k=1}^{N} h(\boldsymbol{x}_j + \boldsymbol{\varepsilon}_k) - y_j \right)^2. \tag{6}$$

Remark 4 *The loss function $L_n(h)$ is slightly different from the loss function used in practice, i.e.,*

$$L_n'(h) = \frac{1}{2n} \sum_{j=1}^{n} \frac{1}{N} \sum_{k=1}^{N} \left(h(\boldsymbol{x}_j + \boldsymbol{\varepsilon}_k) - y_j \right)^2. \tag{7}$$

However, it can be shown that $L_n(h)$ is close to $L'_n(h)$. To see this, note that

$$L'_n(h) - L_n(h) = \frac{1}{2n}\sum_{j=1}^{n}\frac{1}{2N^2}\sum_{k=1}^{N}\sum_{l=1}^{N}\left(h(\boldsymbol{x}_j+\boldsymbol{\varepsilon}_k)-h(\boldsymbol{x}_j+\boldsymbol{\varepsilon}_l)\right)^2. \tag{8}$$

As we will see later in Section 4, we require that the variance of $\boldsymbol{\varepsilon}_k$ to converge to zero, which implies that the right-hand side in (8) *is close to zero.*

In order to minimize (5), we apply the gradient descent method. Since we impose a restriction that the estimator f is in the RKHS $\mathcal{H}_K(\Omega)$, by the representer theorem, it suffices to consider the function space

$$\mathcal{F}_0 = \left\{ f : f(\cdot) = \sum_{j=1}^{n}\sum_{k=1}^{N} w_{jk}K(\cdot - (\boldsymbol{x}_j+\boldsymbol{\varepsilon}_k)), w_{jk}\in\mathbb{R} \right\}.$$

Because the number of parameters in $\mathcal{F}_0$ scales as $n\times N$, which can be prohibitively large if there are too many augmentations, it is often necessary to reduce the flexibility of $\mathcal{F}_0$ in order to minimize the loss function (5). To achieve this, we consider a subspace of $\mathcal{F}_0$, denoted by

$$\mathcal{F} = \left\{ f : f(\cdot) = \sum_{j=1}^{n}\sum_{k=1}^{N} w_{j}K(\cdot - (\boldsymbol{x}_j+\boldsymbol{\varepsilon}_k)), w_{j}\in\mathbb{R} \right\},$$

i.e., all the weights for the different augmented data from the same input $\boldsymbol{x}_j$ are the same. Define an empirical random smoothing kernel function by

$$K_S(\boldsymbol{x}_l - \boldsymbol{x}_j) := \frac{1}{N^2}\sum_{k_1=1}^{N}\sum_{k_2=1}^{N} K(\boldsymbol{x}_l+\boldsymbol{\varepsilon}_{k_1} - (\boldsymbol{x}_j+\boldsymbol{\varepsilon}_{k_2})), \tag{9}$$

whose expectation leads to the following random smoothing kernel function, which plays an important role in the convergence analysis.

Definition 5 (Random smoothing kernel function) *The kernel function K_S defined in* (9) *is the empirical random smoothing kernel function corresponding to the original kernel K. The expectation of K_S with respect to the noise $\boldsymbol{\varepsilon}_k$ is the convoluted kernel function $K * p_\varepsilon$, where $*$ is a convolution operator defined by*

$$(g_1 * g_2)(\boldsymbol{s}) = \int g_1(\boldsymbol{t})g_2(\boldsymbol{s}-\boldsymbol{t})\mathrm{d}\boldsymbol{t},$$

for two functions g_1 and g_2. We call the convoluted kernel function $K * p_\varepsilon$ as the random smoothing kernel function.

Now we can rewrite the loss function $L_n(f)$ in (5) (up to a constant multiplier $1/n$ which is not influenced by the solution of the optimization problem) as

$$L_n(\boldsymbol{w}) = \frac{1}{2} \|\boldsymbol{y} - \mathbf{K}\boldsymbol{w}\|_2^2 , \tag{10}$$

where $\mathbf{K} = (K_S(\boldsymbol{x}_j - \boldsymbol{x}_k))_{jk}$, $\boldsymbol{w} = (w_1, ..., w_n)^T$, and $\boldsymbol{y} = (y_1, ..., y_n)^T$. Following the tradition in Raskutti et al. (2014), consider the gradient descent on the transformed vector $\boldsymbol{\theta} = \sqrt{\mathbf{K}}\boldsymbol{w}$, where the square root can be taken because $\mathbf{K}$ is positive (semi-)definite. Then, we apply gradient descent on the square loss (10) with the transformed vector $\boldsymbol{\theta}$. Initialize $\boldsymbol{\theta}_0 = \boldsymbol{w}_0 = 0$. Taking gradient with respect to $\boldsymbol{\theta}$, direct computation shows that the gradient update is[1]

$$\boldsymbol{\theta}_{\mathfrak{t}} = \boldsymbol{\theta}_{\mathfrak{t}} - \beta_{\mathfrak{t}} \left(\mathbf{K}\boldsymbol{\theta}_{\mathfrak{t}} - \sqrt{\mathbf{K}}\boldsymbol{y}\right), \tag{11}$$

where $\beta_{\mathfrak{t}} > 0$, $\mathfrak{t} = 0, 1, 2, \ldots$ is the learning rate (step size). With parameter $\boldsymbol{w}_{\mathfrak{t}}$ obtained at the $\mathfrak{t}$-th iteration, the corresponding estimator of $f^*(\boldsymbol{x})$ for any point $\boldsymbol{x} \in \Omega$ is defined by

$$f_{\mathfrak{t}}(\boldsymbol{x}) = \boldsymbol{w}_{\mathfrak{t}}^T \mathbf{k}(\boldsymbol{x}), \tag{12}$$

where $\mathbf{k}(\boldsymbol{x}) = (K_S(\boldsymbol{x} - \boldsymbol{x}_1), \ldots, K_S(\boldsymbol{x} - \boldsymbol{x}_n))^T$.

In practice, gradient descent is often paired with weight decay (Krogh and Hertz, 1992) to prevent overfitting and improve generalization (Hu et al., 2021). Therefore, we also consider the gradient descent with weight decay, where the parameter $\boldsymbol{\theta}$ is updated by

$$\boldsymbol{\theta}_{\mathfrak{t}+1} = \boldsymbol{\theta}_{\mathfrak{t}} - \beta_{\mathfrak{t}} \left(\mathbf{K}\boldsymbol{\theta}_{\mathfrak{t}} - \sqrt{\mathbf{K}}\boldsymbol{y}\right) - \alpha_{\mathfrak{t}}\boldsymbol{\theta}_{\mathfrak{t}}, \tag{13}$$

with $\alpha_{\mathfrak{t}} > 0$, $\mathfrak{t} = 0, 1, 2, \ldots$ being the strength of weight decay. The learning rate $\beta_{\mathfrak{t}}$ and weight decay parameter $\alpha_{\mathfrak{t}}$ can vary with $\mathfrak{t}$, but for mathematical convenience, we assume that the step sizes $\beta_{\mathfrak{t}}$ and the weights decay parameter $\alpha_{\mathfrak{t}}$ are not related to the iteration number $\mathfrak{t}$, i.e., $\beta_{\mathfrak{t}} = \beta$ and $\alpha_{\mathfrak{t}} = \alpha$ for all $\mathfrak{t} = 0, 1, 2, \ldots$.

As mentioned in Raskutti et al. (2014), one advantage of early stopping compared with kernel ridge regression is lower computational complexity. Specifically, in kernel ridge regression, one needs to solve a family of quadratic programming problem (or a matrix inversion) for a specified set of regularization parameter, each of which typically requires $O(n^3)$ operations (Caponnetto and Yao, 2010). In early-stopping, the regularization path is given by a sequence of gradient descent update, where each update only involves matrix-vector multiplication of typical $O(n^2)$ operations.

One key difference between the usual early-stopping and our method is that we apply the random smoothing, which introduces extra computation. If there are N augmented data for each $\boldsymbol{x}_j$, then in order to compute the empirical random smoothing kernel, one needs extra $O(n^2N^2)$ operations. Although the proposed random augmentation introduces

1. Although we employ reparameterization as $\boldsymbol{\theta} = \sqrt{\mathbf{K}}\boldsymbol{w}$, the gradient descent can be applied to $\boldsymbol{w}$ directly by $\boldsymbol{w}_{\mathfrak{t}+1} = \boldsymbol{w}_{\mathfrak{t}} - \beta_{\mathfrak{t}} (\mathbf{K}\boldsymbol{w}_{\mathfrak{t}} - \boldsymbol{y}) - \alpha_{\mathfrak{t}}\boldsymbol{w}_{\mathfrak{t}}$, and these two update rules are equivalent.

extra computation, it provides benefits on the theoretical convergence rates and empirical performance, as we will see in Sections 4 and 5.

In this work, we are interested in the prediction error

$$\|f^* - f_{\mathrm{t}}\|_{L_2(P_{\mathbf{X}})}. \tag{14}$$

In the rest of this paper, the following definitions are used. For two positive sequences a_n and b_n, we write $a_n \asymp b_n$ if, for some $C, C' > 0$, $C \leq a_n/b_n \leq C'$. Similarly, we write $a_n \gtrsim b_n$ if $a_n \geq Cb_n$ for some constant $C > 0$, and $a_n \lesssim b_n$ if $a_n \leq C'b_n$ for some constant $C' > 0$. Also, $C, C', c_j, C_j, j \geq 0$ are generic positive constants, of which value can change from line to line.

4. Main Results

In this section, we present our main theoretical results. We begin by collecting all the assumptions that will be used throughout the paper in Section 4.1. Then, in Section 4.2, we consider the case where Ω has a finite intrinsic dimension. Finally, in Section 4.3, we consider the case where $\mathcal{H}(\Omega)$ is a tensor RKHS.

4.1 Assumptions

In this work, we will use the following assumptions.

Assumption 1 *The error ϵ_j's in* (1) *are i.i.d. sub-Gaussian (van de Geer, 2000), i.e., satisfying*

$$C^2(\mathbb{E}e^{|\epsilon_j|^2/C^2} - 1) \leq C', \quad j = 1, ..., n,$$

for some positive constants C and C'.

Assumption 2 *There exists $m_0 > D/2$ such that*

$$c_1(1 + \|\boldsymbol{\omega}\|_2^2)^{-m_0} \leq \mathcal{F}(K)(\boldsymbol{\omega}) \leq c_2(1 + \|\boldsymbol{\omega}\|_2^2)^{-m_0}, \forall \boldsymbol{\omega} \in \mathbb{R}^D.$$

for some positive constants c_1 and c_2.

Assumption 3 (Tensor kernel function) *The kernel function K can be expressed as $K = \prod_{j=1}^D K_j$, where K_j's are one-dimensional kernel functions. There exists $m_0 > 1/2$ such that for $j = 1, \ldots, D$,*

$$c_1(1 + \omega_j^2)^{-m_0} \leq \mathcal{F}(K_j)(\omega_j) \leq c_2(1 + \omega_j^2)^{-m_0}, \forall \omega_j \in \mathbb{R}.$$

for some positive constants c_1 and c_2.

Remark 6 *In this work, we only consider the kernel functions whose Fourier transform has the same orders for the upper and lower bounds. This is because the corresponding RKHS is equivalent to some (tensored) Sobolev space, and it is easier to discuss the relationship between our convergence results with the existing works. It is also possible to analyze the prediction error when the kernel function has Fourier transform has different orders for the upper and lower bounds, but this is mathematically involved and is left for future works.*

Example 1 *A class of kernel functions satisfying Assumption 2 is the isotropic Matérn kernel functions (Williams and Rasmussen, 2006). With reparameterization, the Matérn kernel function is given by*

$$K(\boldsymbol{x}) = \frac{(2\phi\sqrt{m_0 - D/2}\|\boldsymbol{x}\|_2)^{m_0-D/2}}{\Gamma(m_0 - D/2)2^{m_0-D/2-1}} B_{m_0-D/2}(2\phi\sqrt{m_0 - D/2}\|\boldsymbol{x}\|_2), \tag{15}$$

with the Fourier transform (Tuo and Wu, 2016)

$$\mathcal{F}(K)(\boldsymbol{\omega}) = \pi^{-D/2}\frac{\Gamma(m_0)}{\Gamma(m_0 - D/2)}(4\phi^2(m_0 - D/2))^{m_0-D/2}(4\phi^2(m_0 - D/2) + \|\boldsymbol{\omega}\|_2^2)^{-m_0}, \tag{16}$$

where $\phi > 0$, and $B_{m_0-D/2}$ is the modified Bessel function of the second kind. It can be seen that (16) *is bounded above and below by $(1 + \|\boldsymbol{\omega}\|_2^2)^{-m_0}$, up to a constant multiplier.*

Another example satisfying Assumption 2 is the generalized Wendland kernel function (Wendland, 2004; Gneiting, 2002; Chernih and Hubbert, 2014; Bevilacqua et al., 2019; Fasshauer and McCourt, 2015), defined as

$$K_{GW}(\boldsymbol{x}) = \begin{cases} \frac{1}{\mathrm{Beta}(2\kappa,\mu+1)} \int_{\|\phi\boldsymbol{x}\|_2}^1 u(u^2 - \|\phi\boldsymbol{x}\|_2^2)^{\kappa-1}(1-u)^\mu \mathrm{d}u, & 0 \leq \|\boldsymbol{x}\|_2 < \frac{1}{\phi}, \\ 0, & \|\boldsymbol{x}\|_2 \geq \frac{1}{\phi}, \end{cases} \tag{17}$$

where $\phi, \kappa > 0$ and $\mu \geq (D+1)/2 + \kappa$, and Beta *denotes the beta function. Theorem 1 of Bevilacqua et al. (2019) shows that* (17) *satisfies Assumption 2 with $m_0 = (D+1)/2 + \kappa$.*

If the kernel function $K = \prod_{j=1}^D K_j$, and each K_j is a one-dimensional Matérn kernel function or generalized Wendland kernel function, then Assumption 3 is satisfied.

Assumption 4 (Random smoothing noise) *The elements of $\boldsymbol{\varepsilon}_k$ are i.i.d. mean zero sub-Gaussian random variables. σ_n^2's are positive parameters to be specified later in Section 4.*

(C1) (Polynomial noise) There exists $m_\varepsilon > D/2$ such that the characteristic function of $\boldsymbol{\varepsilon}_k$ satisfies

$$c_1(1 + \sigma_n^2\|\boldsymbol{\omega}\|_2^2)^{-m_\varepsilon} \leq \mathbb{E}(e^{i\boldsymbol{\omega}^T\boldsymbol{\varepsilon}_k}) \leq c_2(1 + \sigma_n^2\|\boldsymbol{\omega}\|_2^2)^{-m_\varepsilon}, \forall \boldsymbol{\omega} \in \mathbb{R}^D.$$

(C2) (Tensor Polynomial noise) There exists $m_\varepsilon > 1/2$ such that the characteristic function of $\boldsymbol{\varepsilon}_k$ satisfies

$$c_1\prod_{j=1}^D(1 + \sigma_n^2\omega_j^2)^{-m_\varepsilon} \leq \mathbb{E}(e^{i\boldsymbol{\omega}^T\boldsymbol{\varepsilon}_k}) \leq c_2\prod_{j=1}^D(1 + \sigma_n^2\omega_j^2)^{-m_\varepsilon}, \forall \boldsymbol{\omega} = (\omega_1, \ldots, \omega_D) \in \mathbb{R}^D.$$

(C3) (Gaussian noise) The elements of $\boldsymbol{\varepsilon}_k$ are normally distributed with variance σ_n^2.

Here the constants c_1 and c_2 do not depend on σ_n and m_ε. We call σ_n the smoothing scale in this work.

Example 2 *It is easy to construct distributions satisfying (C1) or (C2). For example, the generalized Laplace distribution with parameter s has a density function (Kozubowski et al., 2013; Kotz et al., 2001)*

$$p_\varepsilon(\boldsymbol{x}) = \frac{2^{1-s}}{(2\pi)^{D/2}\Gamma(s)}(\sqrt{2}\|\boldsymbol{x}\|_2)^{s+D/2}B_{s-D/2}\left(\sqrt{2}\|\boldsymbol{x}\|_2\right), \tag{18}$$

where Γ is the Gamma function, and $B_{s-D/2}$ is the modified Bessel function of the second kind. It can be shown that the generalized Laplace distribution has the characteristic function

$$\mathbb{E}_{\boldsymbol{X}}(e^{i\boldsymbol{\omega}^T\boldsymbol{X}}) = \left(1 + \frac{1}{2}\boldsymbol{\omega}^T\boldsymbol{\omega}\right)^{-s}.$$

Then $\boldsymbol{\varepsilon}_k = \sigma_n\boldsymbol{X}$ satisfies Assumption 4 (C1).

If each component of $\boldsymbol{\varepsilon}_k/\sigma_n$ has a univariate generalized Laplace distribution and all components are independent, then Assumption 4 (C2) is satisfied.

Assumption 1 assumes that the observation error is sub-Gaussian, which is a standard assumption in nonparametric literature. See van de Geer (2000) for example. Assumption 2 assumes that the Fourier transform of the kernel function $K(\cdot - \cdot)$ has an algebraic decay. Under this assumption, Corollary 10.13 of Wendland (2004) shows that the reproducing kernel Hilbert space $\mathcal{H}_K(\mathbb{R}^D)$ coincides with the Sobolev space $\mathcal{W}^{m_0}(\mathbb{R}^D)$, with equivalent norms. More details on this can be found in Section 3.1. Assumption 3 states that the kernel function K has a tensor structure, and the Fourier transform of each component K_j has an algebraic decay. Assumptions 2 and 3 will be used in Sections 4.2 and 4.3, respectively. Assumption 4 imposes conditions on the noise $\boldsymbol{\varepsilon}_k$'s and considers three types of augmentations: polynomial noise, tensor polynomial noise, and Gaussian noise. The corresponding smoothing techniques are referred to as *polynomial smoothing*, *tensor polynomial smoothing*, and *Gaussian smoothing*, respectively.

4.2 Low Intrinsic Dimension Space

We first consider Ω with finite intrinsic dimension. The intrinsic dimension provides a "measure of the complexity" for the region of interest Ω. The definition of the intrinsic dimension depends on the covering number; see Definition 2.1 of van de Geer (2000) for example.

Definition 7 (Covering number) *Consider a subset $\mathcal{A} \subset \mathcal{G}$ where $\mathcal{G}$ is a normed space. For a given $\delta > 0$, the covering number of $\mathcal{A}$, denoted by $\mathcal{N}_{\mathcal{G}}(\delta, \mathcal{A})$, is defined by the smallest integer M such that $\mathcal{A}$ can be covered by M balls with radius δ and centers $\boldsymbol{x}_1, ..., \boldsymbol{x}_M \in \mathcal{G}$.*

Assumption 5 (Low intrinsic dimension) *There exist positive constants c_1 and $d \leq D$ such that for all $\delta \in (0,1)$, we have*

$$\mathcal{N}_{\ell_\infty^D}(\delta, \Omega) \leq c_1 \delta^{-d},$$

where ℓ_∞^D is the $\mathbb{R}^D$ space equipped with ℓ_∞ norm.

For discussion and examples of regions that satisfy Assumption 5, we refer to Hamm and Steinwart (2021a). In particular, if $\Omega \subset \mathbb{R}^D$ is a bounded region with positive Lebesgue measure or a bounded D'-dimensional differentiable manifold, then Assumption 5 holds with $d = D$ and $d = D'$, respectively.

Besides the low intrinsic dimension, our theoretical results depend on the smoothness of the underlying function. Because we are considering function space on a finite intrinsic dimensional space, which may have Lebesgue measure zero, the usual definition of (fractional) Sobolev space via Fourier transform stated in Section 3.1 cannot be directly applied in our case. Thus, we need to introduce our notion of the smoothness assumption for the functions on finite intrinsic dimension space. Specifically, we impose the following assumption on the underlying true function f^*.

Assumption 6 *There exists a region Ω_1 with positive Lebesgue measure and a Lipschitz boundary such that $\Omega \subset \Omega_1$. The underlying true function f^* is well-defined on Ω_1 with $f^* \in \mathcal{W}^{m_f}(\Omega_1)$, where $m_f = \operatorname{argsup}_{m > D/2}\{m : f^* \in \mathcal{W}^m(\Omega_1)\}$, and $m_f > D/2$.*

In Assumption 6, we further assume that the boundary of Ω_1 is "sufficiently regular" (see Leoni, 2017 for the definition of Lipschitz boundary) and Ω can be contained by Ω_1. Thus, the extension theorem (DeVore and Sharpley, 1993) ensures that there exists an extension operator from $L_2(\Omega_1)$ to $L_2(\mathbb{R}^D)$ and the smoothness of each function is maintained. With Assumption 6, we use m_f to denote the smoothness of f^*. By some well-known extension theorems (see, for example, DeVore and Sharpley, 1993; Evans, 2009, Pages 268-272; Stein, 1970, Theorem 5, Page 181), if $D = d$, then our notion of smoothness coincides with the smoothness of functions on the whole space $\mathbb{R}^D$. In addition, we require $f^* \in \mathcal{W}^{m_f}(\Omega_1)$, which implies that $\{m : f^* \in \mathcal{W}^m(\Omega_1)\}$ is a closed interval $[m_f, +\infty)$.

Our notion of low-dimensional region and smoothness is based on the description provided in Section 3 of Hamm and Steinwart (2021a). In Hamm and Steinwart (2021a), a Besov space $B^s_{2,\infty}$ is defined with the same low-dimensional support Ω, using the s-th modulus of smoothness. By the embedding relationship $H^s \subset B^s_{2,\infty}$ (see Page 44 of Edmunds and Triebel, 2008), it can be seen that our definition represents a specific instance of this broader framework.

Now we are ready to present the main theorems in this subsection. Theorems 8 and 9 state the convergence rates when applying polynomial smoothing and Gaussian smoothing, respectively.

Theorem 8 (Polynomial smoothing) *Suppose Assumptions 1, 2, 4 (C1), 5 and 6 are satisfied. Let* $f_t(\boldsymbol{x})$ *be as in* (12) *and* $\beta = n^{-1}C_1$ *with the positive constant* $C_1 \leq (2\sup_{\boldsymbol{x}\in\mathbb{R}^D} K_S(\boldsymbol{x}))^{-1}$. *Suppose the smoothing scale* $\sigma_n \asymp n^{\nu}$ *with* $\nu \leq 0$. *Suppose one of the following holds:*

1. *There is no weight decay in the gradient descent, and the iteration number* t *satisfies* $t \asymp n^{\frac{2(m_0+m_\varepsilon)}{2m_f+d}}\sigma_n^{2m_\varepsilon}$

2. *There is weight decay in the gradient descent with* $\alpha \asymp n^{-1-\frac{2(m_0+m_\varepsilon)}{2m_f+d}}\sigma_n^{-2m_\varepsilon}$, *and the iteration number satisfies* $t \geq C_2(\frac{m_f}{2m_f+d} + 1/2)\log n/(\log(1-\alpha))$ *for some positive constant* C_2.

Then by setting $m_\varepsilon = 2d^{-1}(2D\max(m_0, m_f) + m_0 d)\log n - m_0$ *and*

$$\nu = \begin{cases} -\frac{2(2m_0+2m_\varepsilon)D-(2m_0+2m_\varepsilon-D)d}{(2m_f+d)(4m_\varepsilon D-(2m_0+2(1-(\log n)^{-1})m_\varepsilon-D)d)} < 0, & D > d, \\ 0, & D = d, \end{cases}$$

we have

$$\|f_t - f^*\|^2_{L_2(P_{\mathbf{X}})} = O_{\mathbb{P}}\left(n^{-\frac{2m_f}{2m_f+d}}(\log n)^{2m_f+1}\right).$$

for $N > N_0$, *where* N *is the number of augmentations, and* N_0 *depends on* n *(specified in Equation 43).*

Theorem 9 (Gaussian smoothing) *Suppose Assumptions 1, 2, 4 (C3), 5, and 6 are satisfied. Let* $f_t(\boldsymbol{x})$ *be as in* (12), $\beta = n^{-1}C_1$ *with the positive constant* $C_1 \leq (2\sup_{\boldsymbol{x}\in\mathbb{R}^D} K_S(\boldsymbol{x}))^{-1}$, *and* $\sigma_n \asymp n^{-\frac{1}{2m_f+d}}$. *Suppose one of the following holds:*

1. *There is no weight decay in the gradient descent, and the iteration number* t *satisfies* $t \asymp n^{\frac{2m_0+2m_f}{2m_f+d}}$

2. *There is weight decay in the gradient descent with* $\alpha \asymp n^{-1-\frac{2(m_0+m_\varepsilon)}{2m_f+d}}$, *and the iteration number satisfies* $t \geq C_2(\frac{m_f}{2m_f+d} + 1/2)\log n/(\log(1-\alpha))$ *for some positive constant* C_2.

Then we have

$$\|f^* - \hat{f}_t\|^2_{L_2(P_{\mathbf{X}})} = O_{\mathbb{P}}(n^{-\frac{2m_f}{2m_f+d}}(\log n)^{D+1}), \tag{19}$$

when $N > N_0$, *where* N *is the number of augmentations, and* N_0 *depends on* n *(specified in Equation 73).*

Remark 10 *We require $\beta = n^{-1}C_1$ with the positive constant $C_1 \leq (2\sup_{\boldsymbol{x}\in\mathbb{R}^D} K_S(\boldsymbol{x}))^{-1}$ in both Theorems 8 and 9 is because by Gershgorin's theorem (Varga, 2010), we have for sufficiently large n,*

$$\beta\eta_1(\mathbf{K}) + \alpha \leq \beta n \max_{j,k} |K_S(\boldsymbol{x}_j, \boldsymbol{x}_k)| + \alpha < 1,$$

where $\eta_1(\mathbf{K})$ is the largest eigenvalues of $\mathbf{K}$, which ensures that the gradient descent algorithm can converge.

Remark 11 *In Theorems 8 and 9, the large number of augmentations N_0 is necessary in our theoretical analysis, while in practice, the number of augmentations is usually small. A smaller, but more practical N_0 will be pursued in the future.*

Remark 12 *In this work, we focus exclusively on the performance of the estimator f_t at the final iteration, with a pre-specified iteration number t. While many other studies, such as Yao et al. (2007); Raskutti et al. (2014), use a learning rate that is a function of $\mathfrak{t}$ (typically in the form $\mathfrak{t}^\zeta$ for some $\zeta \in \mathbb{R}$), our decay rate $\alpha_\mathfrak{t}$ and learning rate $\beta_\mathfrak{t}$ are independent of $\mathfrak{t}$ and are chosen to meet the required order as $\mathfrak{t}$ approaches t. Adapting $\alpha_\mathfrak{t}$ and $\beta_\mathfrak{t}$ to $\mathfrak{t}$ could indeed accelerate the training process. However, it is noteworthy that in the case of weight decay, convergence can be achieved in only $\mathcal{O}(\log n)$ iterations, indicating that our current training procedure is also efficient.*

If the region Ω has a positive Lebesgue measure, then it has been shown that the optimal convergence rate is $n^{-m_f/(2m_f+D)}$ (Stone, 1982). By random smoothing, the gradient descent with early stopping can achieve the optimal convergence rate in this case, up to a logarithm term. Furthermore, it can adapt to the low intrinsic dimension case, where Ω can have Lebesgue measure zero. In Hamm and Steinwart (2021a), it is strongly hypothesized that the convergence rate $n^{-m_f/(2m_f+d)}$ is optimal. Although our definition of the smoothness is different, we have the same hypothesis and leave its exploration as a future work.

It is worth noting that our approach differs from that in Hamm and Steinwart (2021a), and therefore, we can investigate the effects of polynomial smoothing, which may have its own interest. Such non-smooth noise can shed light on non-smooth augmentations commonly used in practice. Furthermore, we obtain an identical result as in Hamm and Steinwart (2021a) if we use Gaussian smoothing. Comparing the convergence rates in Theorems 8 and 9, we find that the convergence rate by polynomial smoothing is slightly worse than that of Gaussian smoothing, since $m_f > D/2$ (Assumption 6). In comparison, Eberts and Steinwart (2013) achieved convergence rate of the similar form $n^{-2m_f/(2m_f+d)+\xi}$ by applying kernel ridge regression with Gaussian kernel functions, where ξ can be any value strictly larger than zero. Clearly, this rate is slower than those in Hamm and Steinwart (2021a) and ours. Under additional assumptions such as a compact Riemannian manifold input space and the underlying function having Lipschitz continuity $m_f \in (0, 1]$, Ye and Zhou (2008) derived convergence rates of the form $\left(\log^2(n)/n\right)^{m_f/(8m_f+4d)}$. Instead of kernel ridge regression, Yang and Dunson (2016) focused on Bayesian regression with Gaussian process and proved the convergence rate $n^{-2m_f/(2m_f+d)}(\log n)^{d+1}$. However, their theorem is limited by a compact low dimensional differentiable manifold input space, and the condition $m_f \leq 2$. As a comparison, we do not require such restrictive assumptions.

From a different perspective of early stopping, we consider both cases with and without weight decay, while existing studies only consider the case without weight decay. With weight decay, one can achieve the same convergence rate but with a much smaller iteration number. Specifically, the iteration number should be polynomial in n without weight decay, which can be reduced to polynomial in $\log n$ if one applies weight decay. This also justifies the use of weight decay in practice. Besides, the random smoothing kernel enables us to establish connections with data augmentation and we further explain the effectiveness of using augmentation, which may lead to a new interpretation of using augmentations in deep learning.

Our approach to studying early stopping is distinct from previous studies in the literature (see, e.g., Dieuleveut and Bach, 2016; Yao et al., 2007; Pillaud-Vivien et al., 2018; Raskutti et al., 2014), which typically use integral operator techniques and impose assumptions on the eigenvalues of the kernel function (which always exists by Mercer's theorem). However, such assumptions cannot be easily applied to the low intrinsic dimension case, as it is unclear how eigenvalues behave in this regime. Additionally, previous studies often impose a "source condition" that requires the kernel function to have finite smoothness, which is not satisfied when using Gaussian smoothing to construct the random smoothing kernel. Therefore, even for the special case where the intrinsic dimension is equal to the ambient dimension, Theorems 8 and 9 improve upon previous results in the early stopping literature.

As a special case, it can be shown that trainning a sufficiently overparametrized shallow neural network can be described by a specific kernel called as "neural tangent kernel" (NTK) (Jacot et al., 2018). Chen and Xu (2020) further showed that the NTK induced by the ReLU activation function and Laplace Kernel have the same RKHS. Hence, if we directly choose $m_0 = d/2 + 1/2$, we can see that our convergence results can be applied to the overparametrized shallow neural networks.

4.3 Tensor Reproducing Kernel Hilbert Space

In this section, we consider a low-dimensional structure for the function class, specifically a *tensor reproducing kernel Hilbert space*. Let $K = \prod_{j=1}^{D} K_j$ be kernel functions that satisfy Assumption 3, while Ω can have a low intrinsic dimensional structure, as discussed in Section 4.2, or have a positive Lebesgue measure in $\mathbb{R}^D$.

Our theoretical results in this section are based on mixed smooth Sobolev spaces, denoted by $\mathcal{MW}^m(\mathbb{R}^D)$, where $m > 1/2$. For a function f defined on $\mathbb{R}^D$, the mixed smooth Sobolev norm is defined as

$$\|f\|_{\mathcal{MW}^m(\mathbb{R}^D)} = \left(\int_{\mathbb{R}^D} |\mathcal{F}(f)(\boldsymbol{\omega})|^2 \prod_{j=1}^{D} (1 + |\omega_j|^2)^m \mathrm{d}\boldsymbol{\omega} \right)^{1/2}, \tag{20}$$

and the mixed smooth Sobolev spaces on Ω can be defined via restriction similar to the Sobolev spaces. In fact, the mixed smooth Sobolev space is a tensor product of one-dimensional Sobolev spaces, and it can be shown that $\mathcal{MW}^{m_0}(\mathbb{R}^D)$ is equivalent to the tensor reproducing kernel Hilbert space generated by kernel function $K = \prod_{j=1}^D K_j$ satisfying Assumption 3. Because of such a tensor structure, it is often considered as a reasonable model reducing the complexity in high-dimensional spaces (Kühn et al., 2015; Dũng, 2021). For instance, the mixed smooth Sobolev spaces are utilized in high-dimensional approximation and numerical methods of PDE (Bungartz and Griebel, 1999), data mining (Garcke et al., 2001), and deep neural networks (Dũng, 2021).

If the underlying function belongs to some mixed smooth Sobolev space, then it can be shown that by applying appropriate augmentations, we can achieve a fast convergence rate, which nearly coincides with the minimax rate in the one-dimensional case, up to a logarithmic term. Similar to Assumption 6, we assume that f^* can be extended to some "regular space" with positive Lebesgue measure, as follows.

Assumption 7 *There exists a region Ω_1 with positive Lebesgue measure and a Lipschitz boundary such that $\Omega \subset \Omega_1$, and the underlying true function f^* is well-defined on Ω_1 and $f^* \in \mathcal{MW}^{m_f}(\Omega_1)$.*

The following theorem states the convergence rate when applying tensor polynomial smoothing in the tensor RKHS case.

Theorem 13 (Tensor polynomial smoothing) *Suppose Assumptions 1, 3, 4 (C2), 5, and 7 are satisfied. Let $f_t(\boldsymbol{x})$ be as in (12) and $\beta = n^{-1}C_1$ with the positive constant $C_1 \leq (2\sup_{\boldsymbol{x}\in\mathbb{R}^D} K_S(\boldsymbol{x}))^{-1}$. Let $m_\varepsilon + m_0 \geq m_f$, and the smoothing scale $\sigma_n \asymp 1$.*

Then the following statements are true with $N > N_0$, where N is the number of augmentations, and N_0 depends on n (specified in Equation 83). Suppose one of the following holds:

1. *There is no weight decay in the gradient descent, and the iteration number t satisfies $t \asymp n^{\frac{2(m_0+m_\varepsilon)}{2m_f+1}} (\log n)^{\frac{2(D-1)(m_0+m_\varepsilon)+1}{2m_f+1}}$*
2. *There is weight decay in the gradient descent with $\alpha \asymp n^{-1-\frac{2(m_0+m_\varepsilon)}{2m_f+d}} (\log n)^{\frac{2(D-1)(m_0+m_\varepsilon)+1}{2m_f+1}}$, and the iteration number satisfies $t \geq C_2(\frac{m_f}{2m_f+1} + 1/2)\log n/(\log(1-\alpha))$ for some positive constant C_2.*

Then we have

$$\|f_t - f^*\|^2_{L_2(P_{\mathbf{X}})} = O_{\mathbb{P}}\left(n^{-\frac{2m_f}{2m_f+1}} (\log n)^{\frac{2m_f}{2m_f+1}\left(D-1+\frac{1}{2(m_0+m_\varepsilon)}\right)}\right). \tag{21}$$

Based on Theorem 13, tensor polynomial smoothing leads to a convergence rate of tensor RKHS, which is $O_{\mathbb{P}}(n^{-\frac{2m_f}{2m_f+1}} (\log n)^{\frac{2m_f}{2m_f+1}(D-1+\frac{1}{2(m_0+m_\varepsilon)})})$. This convergence rate is almost the same as the optimal convergence rate in the one-dimensional case $O_{\mathbb{P}}(n^{-\frac{2m_f}{2m_f+1}})$, differing only by a logarithmic term.

Moreover, compared to Theorem 8, Theorem 13 has less stringent requirements for tensor polynomial smoothing when Assumption 7 holds. Specifically, Theorem 13 allows for m_ε to be a constant as long as $m_\varepsilon + m_0 \geq m_f$, whereas Theorem 8 requires m_ε to be comparable to $\log n$. Additionally, while the smoothing scale σ_n in Theorem 8 demands careful selection, Theorem 13 permits a constant smoothing scale σ_n. These differences suggest that the tensor RKHS has a simpler structure than the Sobolev RKHS even in a low intrinsic dimension space. The convergence rate in Theorem 13 does not depend on the low intrinsic dimension of Ω, and is almost dimension-free. Moreover, because the power of the logarithmic term in (21) decreases as m_ε increases, the convergence rate in Theorem 13 decreases as m_ε increases, encouraging the use of a smoother tensor polynomial smoothing for faster convergence. This aligns with the results in Theorem 8 and Theorem 9, as Gaussian smoothing may yield faster convergence rates than polynomial smoothing. Few studies have explored tensor RKHSs with early stopping, and our findings can provide valuable insights into this area.

Remark 14 *For any $\mathcal{W}^{m_f}(\mathbb{R}^D)$ with $m_f > D/2$, an $m^* > 1/2$ can be found for which the embedding relations hold true: $\mathcal{W}^{m_f}(\mathbb{R}^D) \hookrightarrow \mathcal{MW}^{m^*}(\mathbb{R}^D)$ and $\mathcal{MW}^{m^*}(\mathbb{R}^D) \hookrightarrow \mathcal{C}(\mathbb{R}^D)$. Therefore, $\mathcal{W}^{m^*}(\mathbb{R}^D)$ offers a feasible choice as a target space for containing the underlying function, presenting an alternative to the more conventionally used Sobolev spaces.*

Remark 15 *Convolutional Neural Networks (CNNs) can be described using NTKs in the form of tensor products, as shown in Geifman et al. (2022). In their study, Geifman et al. (2022) proved that the NTKs for CNNs are tensor products of kernels whose eigenvalues exhibit polynomial decay. Consequently, by setting $m_0 = \zeta + 2\nu - 3$, where ζ is the number of channels in a CNN and ν depends on the input dimension, we can see that our convergence results can also be applied to CNNs.*

5. Numerical Studies

In this section, we enhance our theoretical findings by experimentally validating the effectiveness of the random smoothing kernel with data augmentation and early stopping on synthetic data sets. We focus on five data spaces with dimensions $D = 1$ ($d = 1$), $D = 2$ ($d = 1, 2$) and $D = 3$ ($d = 1, 2$), as illustrated in Figure 1 and Figure 2, where $\boldsymbol{x}_j$ samples are uniformly drawn.

In our experiments, the underlying function f^* is obtained by drawing random sample paths from the Gaussian process with the Matérn covariance function. This covariance function is widely used in Gaussian process modeling. We adopt the Matérn covariance function with the following form:

$$K_\nu(\boldsymbol{x}) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{\|\boldsymbol{x}\|_2}{\rho} \right)^\nu B_\nu \left(\sqrt{2\nu} \frac{\|\boldsymbol{x}\|_2}{\rho} \right), \tag{22}$$

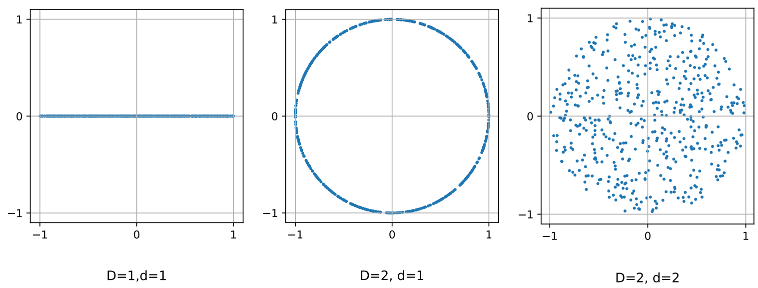

Figure 1: Simulated data spaces in the forms of: line ($D = 1, d = 1$), ring ($D = 2, d = 1$) and disk ($D = 2, d = 2$).

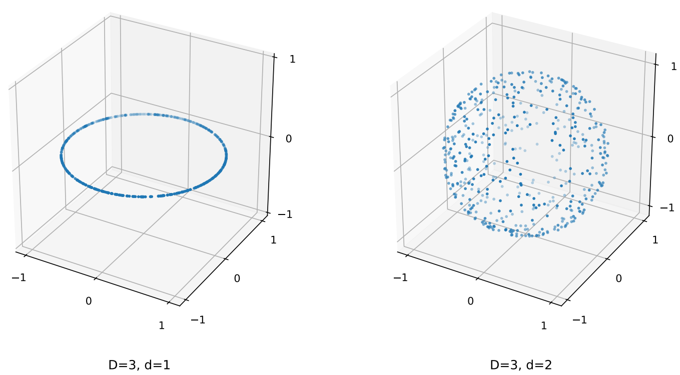

Figure 2: Simulated data spaces in the forms of: ring ($D = 3, d = 1$) and sphere ($D = 3, d = 2$).

where $\sigma, \phi, \nu > 0$, Γ is the Gamma function, and B_ν is the modified Bessel function of the second kind. In order to make f^* smoother, we set the smoothness parameter $\nu = 5.0$ for Matérn kernel (22). The error ϵ_j's are i.i.d. Gaussian with mean zero and variance 0.01.

We utilize two-hidden-layer neural networks with ReLU activation (Nair and Hinton, 2010) as our predictor. Each hidden layer of the neural network comprises 100 nodes, and all weights are initialized using Kaiming Initialization (He et al., 2015). For random smoothing, we experiment with both non-smooth Laplace noise and smooth Gaussian noise. To be precise, each element of $\boldsymbol{\varepsilon}_k$ is randomly sampled from either $\mathcal{N}(0, \sigma^2)$ or $Laplace(0, b)$. For more experiment details and additional results, we refer to Appendix N.

Figure 3 presents a visualization of the underlying truth (blue curve), training data (blue dots), and neural network predictions (orange dots) when the training size is 50. The underlying truth is smooth since we use a smooth kernel. However, the neural network predictions without random smoothing are not smooth due to the low smoothness of the ReLU activation function and tend to overfit the noise. Upon applying random smoothing, the neural network predictions become smoother and approach the underlying truth.

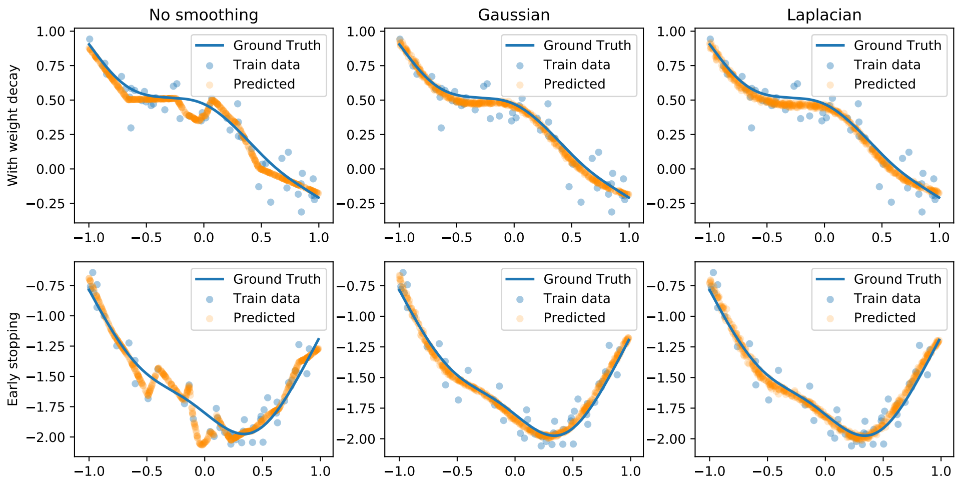

Figure 3: Visualization of the underlying truth (blue curve), training data (blue dots), and neural network predictions (orange dots) when training size is 50, where the first and second rows represent cases with weight decay and early stopping, respectively. It is obvious to see that the optimization without random smoothing will be more vulnerable to noise.

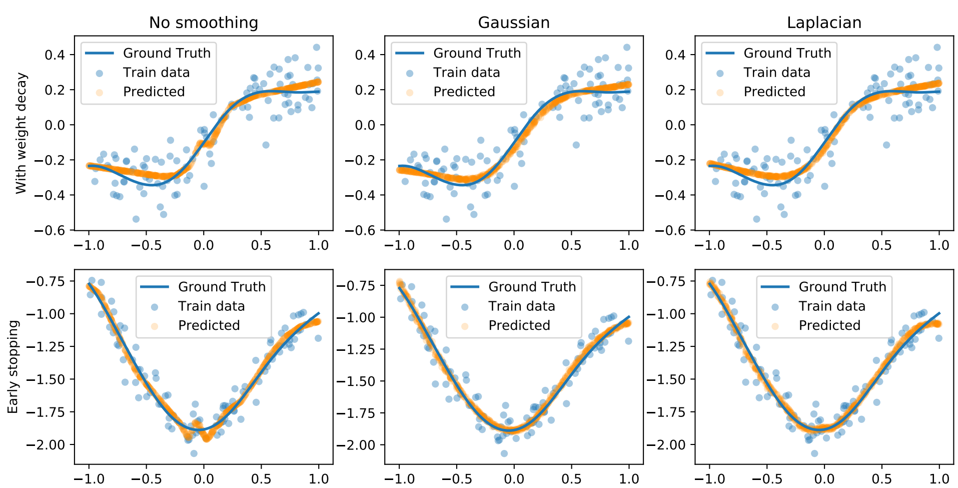

Figure 4: Underlying truth (blue curve), training data (blue dots), and neural network predictions (orange dots) when training size is 100.

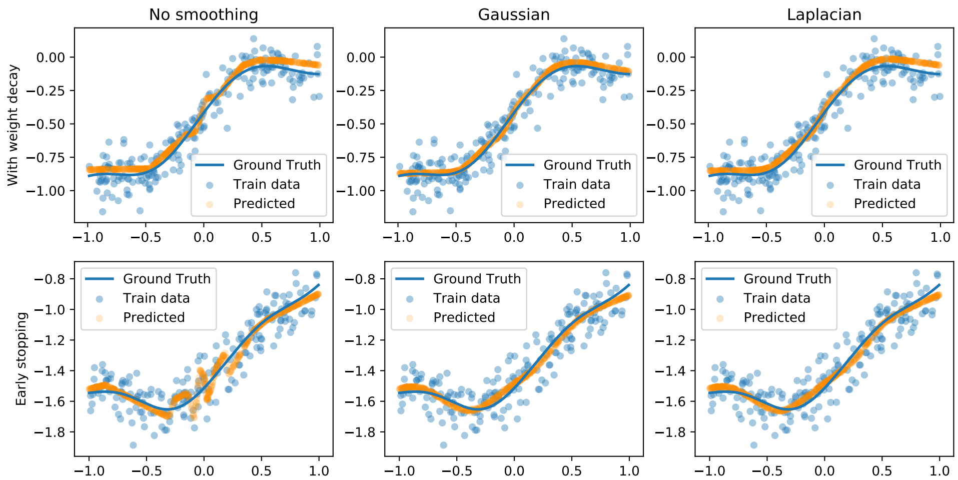

Figure 5: Underlying truth (blue curve), training data (blue dots), and neural network predictions (orange dots) when training size is 200.

Figure 4 and Figure 5 further show the underlying truth (blue curve), training data (blue dots), and neural network predictions (orange dots) when the training size is 100 and 200, respectively. Although increasing the training size improves smoothness in cases like size 200 with weight decay, the fitted curve still experiences a perturbation from overfitted noise compared to examples where random smoothing is applied.

Table 1 presents a summary of the test l_2 loss under different settings. Both Gaussian smoothing and polynomial smoothing (random smoothing with Laplacian noise) improve the l_2 loss in all settings, demonstrating the effectiveness of random smoothing. Figure 6 further investigates how the l_2 loss changes concerning the smoothing scale σ_n when $D = 1$. The plot shows a U-shaped curve, indicating that an optimal smoothing can minimize the l_2 loss, while either smaller or larger values will result in a larger l_2 loss. It is worth noting that when the training size is small, such as size 50, the U-shape curve in Figure 6 may be less distinct due to noise introduced by early stopping based on a small validation set. Another observation from Figure 6 is that the optimal smoothing scales exhibit a decreasing trend as the sample size increases, as indicated by Theorem 8 and Theorem 9. Additionally, Figures 7-8 and Figures 9-10 depict the U-shaped curves of l_2 loss changes concerning smoothing scale when $D = 2$ $(d = 1, 2)$ and $D = 3$ $(d = 1, 2)$, respectively. While it is possible that some red points may not be accurately placed due to a small validation set, the optimal smoothing scales exhibit a decreasing trend with respect to training size, which is consistent with the trend observed in $D = 1$ as depicted in Figure 6.

Dim	Type	With weight decay			Early stopping		
		Training size			Training size		
		50	100	200	50	100	200
D=1,d=1	G	$\underline{\text{1.6765e-03}}$	$\underline{\text{9.3367e-04}}$	$\underline{\text{8.2806e-04}}$	$\underline{\text{1.3468e-03}}$	$\underline{\text{7.5579e-04}}$	$\underline{\text{5.8775e-04}}$
	L	1.7466e-03	9.8343e-04	9.1924e-04	2.0638e-03	9.2128e-04	6.5118e-04
	N	1.9381e-03	1.3045e-03	1.1135e-03	2.2168e-03	1.2985e-03	8.4292e-04
D=2,d=1	G	$\underline{\text{4.1084e-03}}$	$\underline{\text{1.9663e-03}}$	1.8606e-03	$\underline{\text{2.9608e-03}}$	$\underline{\text{1.3643e-03}}$	$\underline{\text{9.5987e-04}}$
	L	4.7216e-03	2.1707e-03	$\underline{\text{1.7646e-03}}$	3.0796e-03	1.5640e-03	1.0766e-03
	N	7.8836e-03	3.4316e-03	2.6008e-03	5.1484e-03	2.6596e-03	1.6057e-03
D=2,d=2	G	6.4676e-03	$\underline{\text{2.9491e-03}}$	2.2136e-03	$\underline{\text{6.7205e-03}}$	$\underline{\text{3.5027e-03}}$	$\underline{\text{1.7132e-03}}$
	L	$\underline{\text{6.4208e-03}}$	3.1423e-03	$\underline{\text{2.1842e-03}}$	8.2725e-03	3.9418e-03	1.7674e-03
	N	9.2474e-03	4.5782e-03	2.5810e-03	1.2628e-02	6.2301e-03	3.1396e-03
D=3,d=1	G	$\underline{\text{4.0382e-03}}$	$\underline{\text{1.9133e-03}}$	$\underline{\text{1.3327e-03}}$	$\underline{\text{1.7104e-02}}$	$\underline{\text{6.8696e-03}}$	$\underline{\text{3.7194e-03}}$
	L	4.8212e-03	2.1033e-03	1.9527e-03	1.7297e-02	7.0159e-03	3.7916e-03
	N	7.4013e-03	3.4172e-03	1.9693e-03	2.3458e-02	8.8306e-03	5.1156e-03
D=3,d=2	G	1.6599e-02	$\underline{\text{6.9336e-03}}$	4.4334e-03	$\underline{\text{1.4852e-02}}$	7.1306e-03	$\underline{\text{3.7147e-03}}$
	L	$\underline{\text{1.6498e-02}}$	7.2578e-03	$\underline{\text{3.9938e-03}}$	1.5167e-02	$\underline{\text{6.6471e-03}}$	3.8615e-03
	N	2.0987e-02	8.1158e-03	4.5752e-03	2.0178e-02	8.4932e-03	4.9460e-03

Table 1: Test l_2 loss of SGD with early stopping. “G”, “L”, and “N” correspond to random smoothing with Gaussian noise, random smoothing with Laplacian noise, and no random smoothing. The smallest losses are underlined.

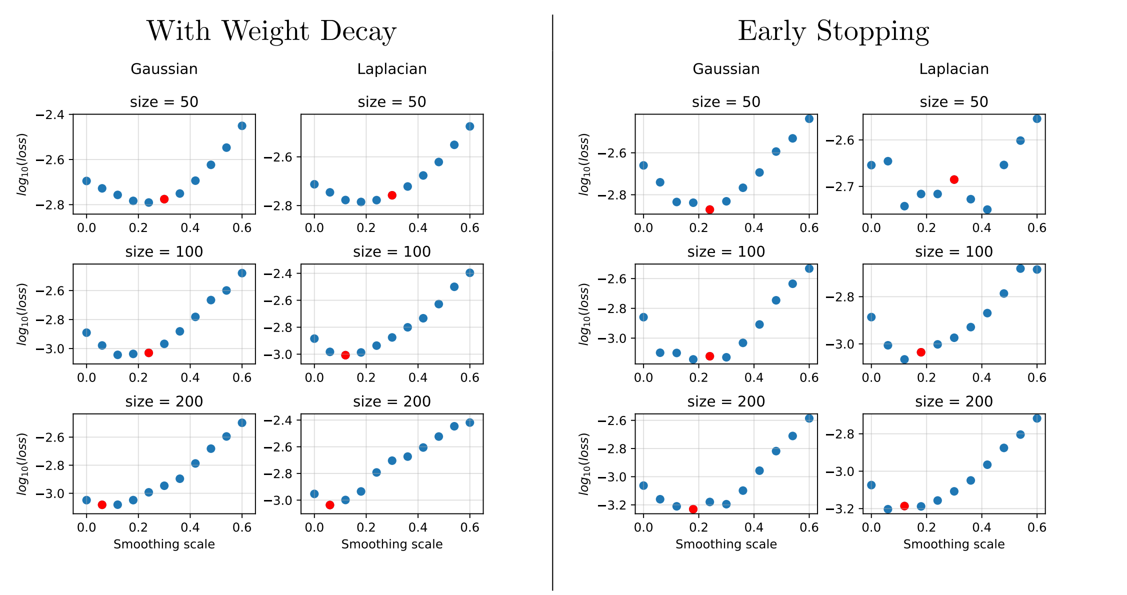

Figure 6: Loss changes according to smoothing scale with training size increase from 50 to 200 in the data space of $D = 1$, $d = 1$. The red points represent the optimal smoothing scales selected based on the validation set.

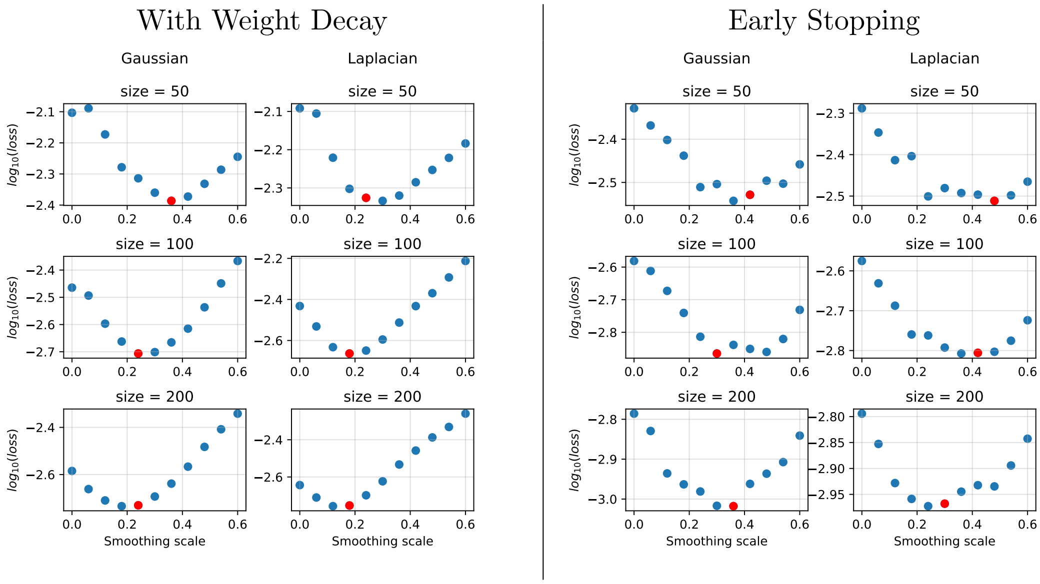

Figure 7: Loss changes according to smoothing scale with training size increase from 50 to 200 in the data space of $D = 2$, $d = 1$. The red points represent the optimal smoothing scales selected based on the validation set.

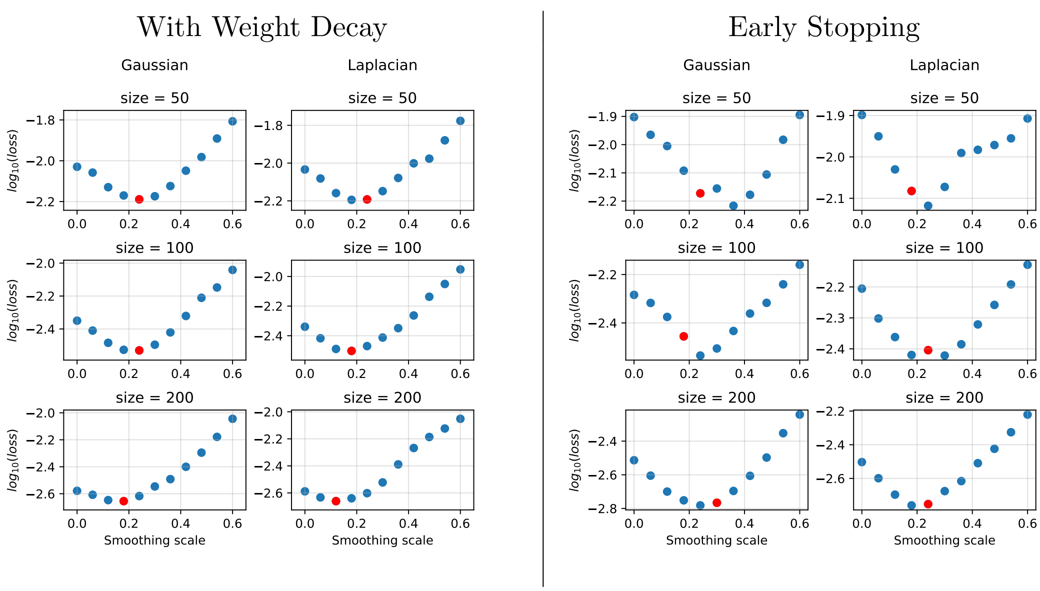

Figure 8: Loss changes according to smoothing scale with training size increase from 50 to 200 in the data space of $D = 2$, $d = 2$. The red points represent the optimal smoothing scales selected based on the validation set.

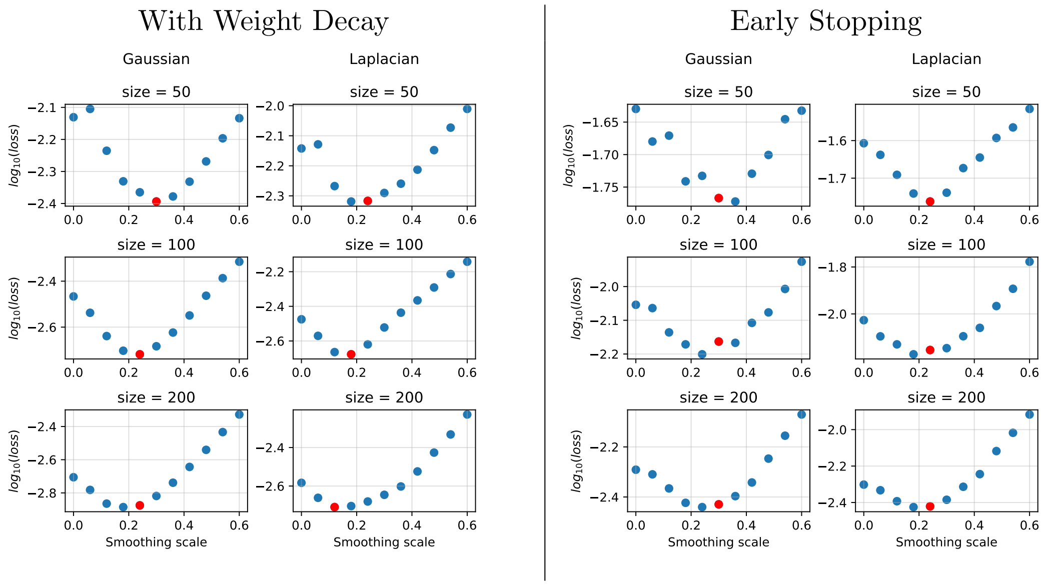

Figure 9: Loss changes according to smoothing scale with training size increase from 50 to 200 in the data space of $D = 3$, $d = 1$. The red points represent the optimal smoothing scales selected based on the validation set.

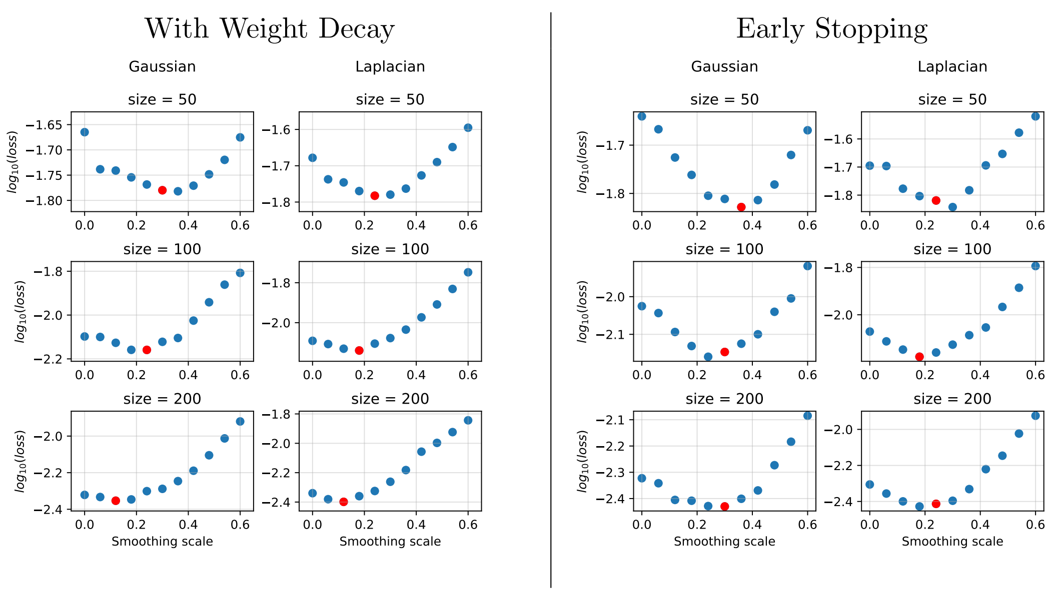

Figure 10: Loss changes according to smoothing scale with training size increase from 50 to 200 in the data space of $D = 3$, $d = 2$. The red points represent the optimal smoothing scales selected based on the validation set.

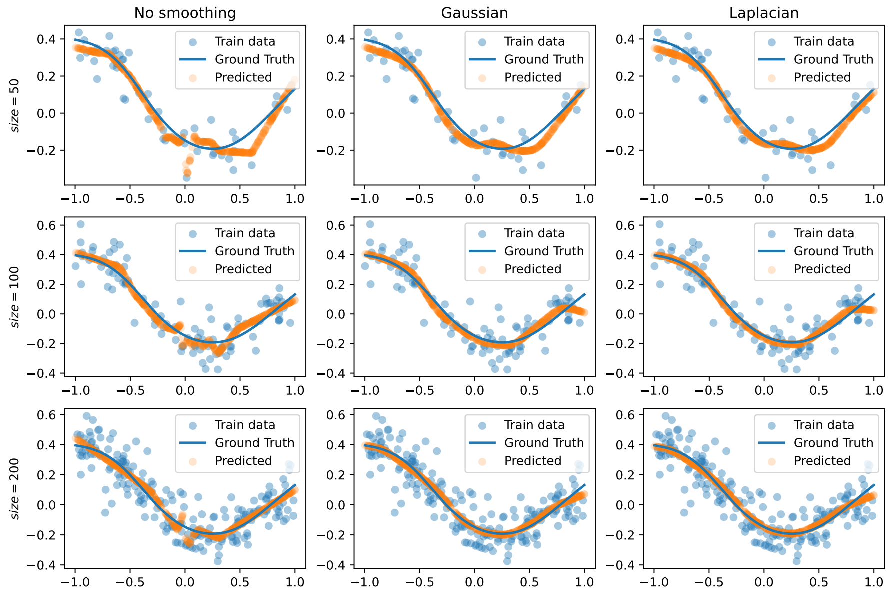

Figure 11: Visualization of the underlying truth (blue curve), training data (blue dots), and neural network predictions (orange dots) when the loss function L'_n is applied. Different rows represent different training sizes.

5.1 Comparison under Different Loss Functions

To illustrate that random smoothing can achieve a similar improvement in the performance of the loss function L'_n in (7), which is slightly different from the one we analyze (L_n in Equation 6), we conducted additional experiments focusing on early stopping with a dimension of $D = 1(d = 1)$. As the performance of L'_n is unstable, we set the number of augmented samples $N = 5000$ and the region for the smoothing scale is from 0 to 0.003. The remaining experimental setup is the same as in L_n.

Figure 11 presents a visualization of the fitted curve with different training sizes. We take the average of the estimator instead of utilizing the prediction directly. The performance is similar to that of L_n (in Figure 3-5), where optimization without random smoothing is more vulnerable to noise, although an increased training size can improve smoothness.

Table 2 summarizes the test l_2 loss with different settings. Both L_n and L'_n in different training sizes can be improved with random smoothing. However, the test loss of L'_n is slightly higher compared to L_n. Figure 12 further demonstrates the varying losses according to the smoothing scale with different loss functions. Although a U-shaped curve can be obtained by L'_n, the optimal smoothing scale is inconsistent with that of L_n, which decreases as the training size increases.

Type	L'_n			L_n		
	Training size			Training size		
	50	100	200	50	100	200
G	<u>2.0824e-03</u>	1.1469e-03	7.7571e-04	<u>1.3468e-03</u>	<u>7.5579e-04</u>	<u>5.8775e-04</u>
L	2.3245e-03	<u>1.1320e-03</u>	<u>7.0490e-04</u>	2.0638e-03	9.2128e-04	6.5118e-04
N	3.5092e-03	1.4109e-03	7.9209e-04	2.2168e-03	1.2985e-03	8.4292e-04

Table 2: Test l_2 loss of SGD with early stopping. We focus on the comparison between different loss functions L'_n and L_n with dimension $D = 1(d = 1)$. "G", "L", and "N" correspond to random smoothing with Gaussian noise, random smoothing with Laplacian noise, and no random smoothing. The smallest losses are underlined.

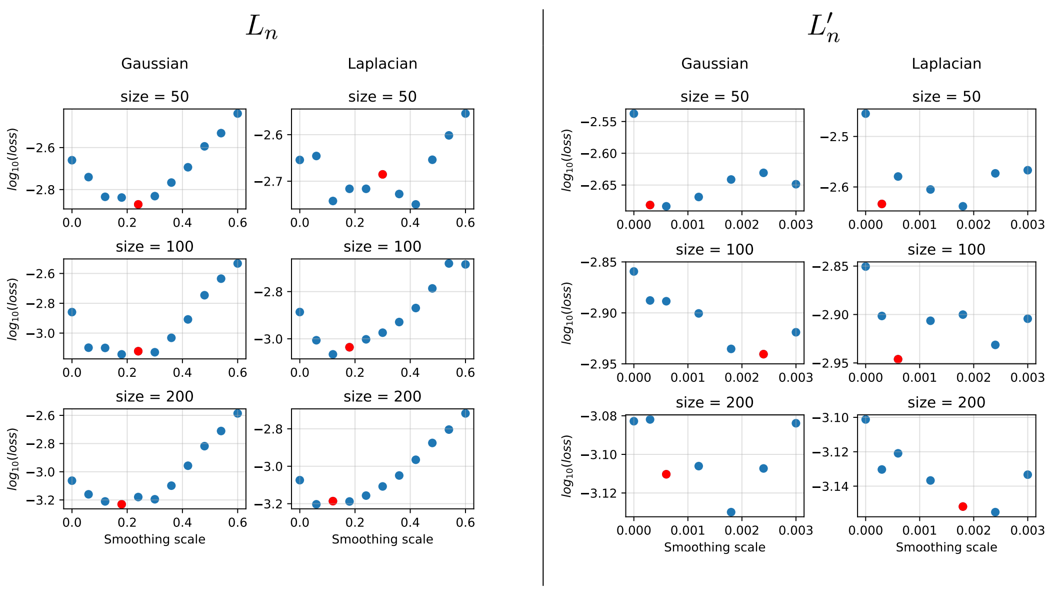

Figure 12: Comparison of different loss changes according to smoothing scale with training size increase from 50 to 200 in the data space of $D = 1$, $d = 1$. The red points represent the optimal smoothing scales selected based on the validation set.

Data set	Type	With weight decay					Early stopping				
		Training size					Training size				
		25	50	100	200	400	25	50	100	200	400
Iris	G	93.33	94.67	-	-	-	$\underline{92.11}$	$\underline{94.56}$	-	-	-
	L	$\underline{94.67}$	$\underline{97.33}$	-	-	-	91.16	94.01	-	-	-
	N	82.67	90.67	-	-	-	88.16	93.88	-	-	-
Rice	G	$\underline{89.50}$	$\underline{88.77}$	$\underline{91.18}$	$\underline{91.86}$	92.44	87.80	$\underline{88.74}$	$\underline{90.17}$	$\underline{90.78}$	$\underline{91.81}$
	L	89.19	88.08	90.87	91.23	$\underline{92.49}$	$\underline{88.99}$	88.57	89.57	90.73	91.33
	N	88.19	87.98	90.18	90.29	92.23	88.38	88.15	89.04	89.94	90.49
Dry Bean	G	73.83	$\underline{82.94}$	86.87	90.04	91.32	$\underline{75.64}$	$\underline{82.93}$	$\underline{86.51}$	88.17	89.86
	L	$\underline{74.96}$	81.97	$\underline{88.14}$	$\underline{90.35}$	$\underline{91.69}$	75.76	82.48	86.43	$\underline{88.76}$	$\underline{90.00}$
	N	73.32	80.38	86.04	88.94	91.31	74.18	81.12	86.35	88.23	89.76
Raisin	G	76.67	82.00	$\underline{85.78}$	$\underline{86.67}$	$\underline{85.78}$	79.27	$\underline{81.93}$	81.25	83.63	$\underline{86.17}$
	L	$\underline{80.00}$	$\underline{83.56}$	85.56	86.22	85.56	78.71	81.79	81.50	$\underline{83.99}$	85.80
	N	77.33	83.33	84.67	85.56	85.33	$\underline{79.84}$	81.38	$\underline{81.56}$	82.68	84.38

Table 3: Test accuracy of different real world data set. "G", "L", and "N" correspond to random smoothing with Gaussian noise, random smoothing with Laplacian noise, and no random smoothing. The highest accuracies are underlined.

6. Experiments on Real-world Data set

To demonstrate the practical application of our theoretical findings, we conducted classification tasks on four real-world data sets: Iris (Fisher, 1988), Rice (Cammeo and Osmancik) (mis, 2019), Dry Bean (mis, 2020), and Raisin (Çinar et al., 2023).

Following the experiments in Section 5, We use a two-hidden-layer neural network with $N = 1000$ augmented samples and replace the l_2 loss with Cross Entropy loss. To isolate the influence of random smoothing, we apply a constant weight decay strength for each data set. For early stopping without weight decay, we evaluate the validation set every 200 steps and select the highest accuracy. We conduct grid searches to determine the optimal smoothing scale for each data set and repeat the experiment 5 times to report the average accuracy on the test set.

Table 3 presents the test classification accuracy of four real-world data sets with varying training sizes. Due to the sample size limitation, we only consider 25 and 50 training data for Iris. Almost all settings show a significant improvement in test accuracy after applying the random smoothing method, especially for smaller data sets like Iris.

7. Conclusions and Discussion

This work studies random smoothing kernel and random smoothing regularization, which have a natural relationship with data augmentations. We consider two cases: when the region Ω has a low intrinsic dimension, or when the kernel function can be presented as a product of one-dimensional kernel functions. In both cases, we show that by applying random smoothing, with appropriate early stopping and/or weight decay techniques, the resulting estimator can achieve fast convergence rates, regardless of the kernel function used in the construction of the random smoothing kernel estimator.

There are several directions that could be pursued in future research. First, while we consider noise injection to construct augmentations and use non-smooth noise to interpret practical non-smooth augmentation techniques, such as random crop, random mask, and random flip, this interpretation may not be perfect. For example, the behavior of adding noise may differ from that of random crop. Furthermore, these practical techniques may also introduce some prior knowledge on the geometry of the low intrinsic dimension. A sharper characterization of practical augmentation techniques is needed and will be pursued in future work.

Second, while we consider gradient descent, we believe that our results can be generalized to the stochastic gradient descent method. However, the discussion of the latter is beyond the scope of the current work.

Third, we mainly consider regression in this work, where the square loss is a natural choice. An interesting extension is to study whether the results remain true when considering classification, which requires the study of other loss functions, such as cross-entropy loss and hinge loss.

Acknowledgments

The authors are grateful to the AE and reviewers for their very constructive comments and suggestions. The authors also thank Ahri Li for his helpful suggestions and this paper is also in memory of him. Wang's work was supported by NSFC Grant 12101149. This work was supported in part by NSFC/RGC Joint Research Scheme Grant N_HKUST635/20 and HKRGC Grant 16308321.

Appendix A. Analysis of Gradient Update and Error Decomposition

Let $\mathbf{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$, $\alpha > 0$ if there is weight decay, and $\alpha = 0$ if there is no weight decay. By the gradient update rule, we have

$$\begin{aligned} f_{\mathrm{t}}(\mathbf{X}) =& \mathbf{K}\boldsymbol{w}_{\mathrm{t}} = \sqrt{\mathbf{K}}\boldsymbol{\theta}_{\mathrm{t}} \\ =& \sqrt{\mathbf{K}}\boldsymbol{\theta}_{\mathrm{t}} - \beta\sqrt{\mathbf{K}}\left(\mathbf{K}\boldsymbol{\theta}_{\mathrm{t}} - \sqrt{\mathbf{K}}\boldsymbol{y}\right) - \alpha\sqrt{\mathbf{K}}\boldsymbol{\theta}_{\mathrm{t}} \\ =& ((1-\alpha)\mathbf{I} - \beta\mathbf{K})f_{\mathrm{t}}(\mathbf{X}) + \beta\mathbf{K}\boldsymbol{y}, \end{aligned}$$

which implies

$$\begin{aligned} & f_{\mathrm{t}+1}(\mathbf{X}) - \beta(\alpha\mathbf{I} + \beta\mathbf{K})^{-1}\mathbf{K}\boldsymbol{y} = ((1-\alpha)\mathbf{I} - \beta\mathbf{K})(f_{\mathrm{t}}(\mathbf{X}) - \beta(\alpha\mathbf{I} + \beta\mathbf{K})^{-1}\mathbf{K}\boldsymbol{y}) \\ =& \ldots = -((1-\alpha)\mathbf{I} - \beta\mathbf{K})^{\mathrm{t}+1}\beta(\alpha\mathbf{I} + \beta\mathbf{K})^{-1}\mathbf{K}\boldsymbol{y}, \end{aligned} \tag{23}$$

where we recall $f_0(\mathbf{X}) = \mathbf{0}$. If there is weight decay (i.e., $\alpha > 0$), then it can be seen that

$$f_{\mathrm{t}+1}(\mathbf{X}) - \mathbf{K}(\alpha/\beta\mathbf{I} + \mathbf{K})^{-1}\boldsymbol{y} = -((1-\alpha)\mathbf{I} - \beta\mathbf{K})^{\mathrm{t}+1}\beta(\alpha\mathbf{I} + \beta\mathbf{K})^{-1}\mathbf{K}\boldsymbol{y}. \tag{24}$$

If there is no weight decay (i.e., $\alpha = 0$), then by rearrangement of (23), we obtain

$$f_{\mathrm{t}+1}(\mathbf{X}) = \left(\mathbf{I} - (\mathbf{I} - \beta\mathbf{K})^{\mathrm{t}+1}\right)\boldsymbol{y}. \tag{25}$$

The estimator after t-th iteration can be obtained by

$$f_{\mathrm{t}}(\boldsymbol{x}) = \boldsymbol{w}_{\mathrm{t}}^T\mathbf{k}(\boldsymbol{x}) = \mathbf{k}(\boldsymbol{x})^T\mathbf{K}^{-1}f_{\mathrm{t}}(\mathbf{X}). \tag{26}$$

Note that the kernel matrix $\mathbf{K}$ is generated by the empirical kernel K_S defined in (9). By taking the expectation with respect to $\boldsymbol{\varepsilon}_{k_1}$ and $\boldsymbol{\varepsilon}_{k_2}$, we define the expected smoothing kernel $\tilde{K}_S$ as

$$\tilde{K}_S(\boldsymbol{x}, \boldsymbol{x}') = \int_{\mathbb{R}^D}\int_{\mathbb{R}^D} K(\boldsymbol{x} + \boldsymbol{\epsilon} - (\boldsymbol{x}' + \boldsymbol{\epsilon}'))p_\varepsilon(\boldsymbol{\epsilon})p_\varepsilon(\boldsymbol{\epsilon}')\mathrm{d}\boldsymbol{\epsilon}\mathrm{d}\boldsymbol{\epsilon}'. \tag{27}$$

Since $\tilde{K}_S$ is close to the empirical version of the smoothing kernel K_S, we can consider the gradient flow with respect to the kernel function $\tilde{K}_S$. The error analysis between $\tilde{K}_S$ and K_S is provided in Appendix B.

Let g_{t} be the function obtained at t-th iteration by the gradient update rule with respect to the kernel function $\tilde{K}_S$. Analogous to (24) and (25), we have

$$g_{\mathrm{t}}(\mathbf{X}) = \tilde{\mathbf{K}}(\alpha/\beta\mathbf{I} + \tilde{\mathbf{K}})^{-1}\boldsymbol{y} - ((1-\alpha)\mathbf{I} - \beta\tilde{\mathbf{K}})^{\mathrm{t}}\beta(\alpha\mathbf{I} + \beta\tilde{\mathbf{K}})^{-1}\mathbf{K}\boldsymbol{y}, \tag{28}$$

if there is weight decay, and

$$g_{\mathrm{t}}(\mathbf{X}) = \left(\mathbf{I} - (\mathbf{I} - \beta\tilde{\mathbf{K}})^{\mathrm{t}}\right)\boldsymbol{y}, \tag{29}$$

if there is no weight decay, where $\tilde{\mathbf{K}} = (\tilde{K}_S(\boldsymbol{x}_j - \boldsymbol{x}_k))_{jk}$. Similarly, the predictor of $f^*(\boldsymbol{x})$ using the kernel function $\tilde{K}$ can be obtained by

$$g_{\mathrm{t}}(\boldsymbol{x}) = \tilde{\mathbf{k}}(\boldsymbol{x})^T\tilde{\mathbf{K}}^{-1}g_{\mathrm{t}}(\mathbf{X}). \tag{30}$$

Thus, the empirical error $\|f_{\mathrm{t}}(\mathbf{X}) - f^*(\mathbf{X})\|_2$ can be decomposed by

$$\|f_{\mathrm{t}}(\mathbf{X}) - f^*(\mathbf{X})\|_2 \leq \|f_{\mathrm{t}}(\mathbf{X}) - g_{\mathrm{t}}(\mathbf{X})\|_2 + \|g_{\mathrm{t}}(\mathbf{X}) - f^*(\mathbf{X})\|_2. \tag{31}$$

Appendix B. Error of Data Augmentation

We first consider bounding the difference between the empirical smoothing kernel function

$$K_S(\boldsymbol{x} - \boldsymbol{x}') = \frac{1}{N^2}\sum_{k=1}^N\sum_{j=1}^N K(\boldsymbol{x} + \varepsilon_j - (\boldsymbol{x}' + \varepsilon_k)),$$

and the expected smoothing kernel function

$$\tilde{K}_S(\boldsymbol{x} - \boldsymbol{x}') = \mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}'}\left(K(\boldsymbol{x} + \boldsymbol{\varepsilon} - (\boldsymbol{x}' + \boldsymbol{\varepsilon}'))\right) = \int_{\mathbb{R}^D}\int_{\mathbb{R}^D} K(\boldsymbol{x} + \boldsymbol{\varepsilon} - (\boldsymbol{x}' + \boldsymbol{\varepsilon}'))p_\varepsilon(\boldsymbol{\varepsilon})p_\varepsilon(\boldsymbol{\varepsilon}')\mathrm{d}\boldsymbol{\varepsilon}\mathrm{d}\boldsymbol{\varepsilon}'.$$

Specifically, we have the following lemma.

Lemma 16 *If Assumption 2 or 3, and Assumption 4 are satisfied , then*

$$\sup_{\boldsymbol{x},\boldsymbol{x}'\in\Omega}\left|\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}'}\big(K(\boldsymbol{x}+\boldsymbol{\varepsilon}-(\boldsymbol{x}'+\boldsymbol{\varepsilon}'))\big)-\frac{1}{N^2}\sum_{k=1}^{N}\sum_{j=1}^{N}K(\boldsymbol{x}+\boldsymbol{\varepsilon}_j-(\boldsymbol{x}'+\boldsymbol{\varepsilon}_k))\right|=O_{\mathbb{P}}\left(\sqrt{\frac{\log N}{N}}\right).$$

Based on Lemma 16, we can obtain an upper bound of $\|f_{\mathrm{t}}-g_{\mathrm{t}}\|_{L_\infty(\Omega)}$ as follows. Recall that $\mathbf{K}=(K_S(\boldsymbol{x}_j-\boldsymbol{x}_k))_{j,k=1}^n$, $\tilde{\mathbf{K}}=(\tilde{K}_S(\boldsymbol{x}_j-\boldsymbol{x}_k))_{j,k=1}^n$. Let $\eta_1(\mathbf{K})$ and $\eta_n(\mathbf{K})$ be the largest and smallest eigenvalues of $\mathbf{K}$, respectively. Let $\eta_n(\tilde{\mathbf{K}})$ be the smallest eigenvalue of $\tilde{\mathbf{K}}$.

Lemma 17 *Suppose Assumption 2 or 3, and Assumption 4 are satisfied. Furthermore, assume that*

$$\frac{1}{2}\eta_n(\tilde{\mathbf{K}})\geq n\sqrt{\frac{\log N}{N}},\tag{32}$$

and the learning rate β satisfies $\beta\eta_1(\mathbf{K})+\alpha<1$, where $\alpha=0$ if there is no weight decay, and $\alpha>0$ if there is weight decay. Then we have

$$\sup_{\mathrm{t}\geq 1}\|f_{\mathrm{t}}-g_{\mathrm{t}}\|_{L_\infty(\Omega)}=O_{\mathbb{P}}\left(\frac{n^2\sqrt{\log N/N}}{\eta_n(\tilde{\mathbf{K}})^2}\right),$$

where the probability is with respect to the augmentation $\boldsymbol{\varepsilon}$.

Since $\tilde{\mathbf{K}}$ and $\eta_n(\tilde{\mathbf{K}})$ are determined by the data $(\boldsymbol{x}_j,y_j)$, $j=1,...,n$, the left-hand side of (32) is not depending on N. Therefore, the condition (32) can be fulfilled if we add sufficient augmentations. In the next lemma, we provide a more explicit lower bound of $\eta_n(\tilde{\mathbf{K}})$ in (32) in terms of $\boldsymbol{x}_j$'s.

Lemma 18 *Let $q_{\mathbf{X}}$ be the separation distance defined as*

$$q_{\mathbf{X}}=\frac{1}{2}\min_{j\neq k}\|\boldsymbol{x}_j-\boldsymbol{x}_k\|_2.$$

The minimum eigenvalue of $\tilde{\mathbf{K}}$, denoted by $\eta_n(\tilde{\mathbf{K}})$, is lower bounded as follows.

1. *if Assumption 2 and Assumption 4 (C1) are satisfied, then*
$$\eta_n(\tilde{\mathbf{K}})\geq C_1\big(1+4M^2\big)^{-m_0}\big(1+4\sigma_n^2M^2\big)^{-m_\varepsilon}M^D;$$
2. *if Assumption 3 and Assumption 4 (C2) are satisfied, then*
$$\eta_n(\tilde{\mathbf{K}})\geq C_2\big(1+4M^2\big)^{-m_0D}\big(1+4\sigma_n^2M^2\big)^{-m_\varepsilon D}M^D;$$
3. *if Assumption 2 and Assumption 4 (C3) are satisfied, then*
$$\eta_n(\tilde{\mathbf{K}})\geq C_3\big(1+4M^2\big)^{-m_0}e^{-8\sigma_n^2M^2}M^D,$$

where C_i's are constants only depending on D, $M=\frac{12}{q_{\mathbf{X}}}\big(\frac{\pi\Gamma^2(\frac{D}{2}+1)}{9}\big)^{\frac{1}{D+1}}$, and $\Gamma(\cdot)$ denotes the Gamma function.

The proofs of the above three lemmas are put in Appendix H.

Appendix C. A Comparison Theorem

In this section, we provide a byproduct, which is a generic comparison theorem between the early-stopping without weight decay and the kernel ridge regression estimator. Let K_1 be a positive definite kernel function. The kernel ridge regression is defined by

$$\tilde{g} = \operatorname*{argmin}_{f \in \mathcal{H}_{K_1}(\Omega)} \|f - \boldsymbol{y}\|_n^2 + \lambda \|f\|^2_{\mathcal{H}_{K_1}(\Omega)}, \tag{33}$$

where $\boldsymbol{y} = (y_1, ..., y_n)^T$, y_j's are as in (1), and $\lambda > 0$ is a regularization parameter. The main theorem in this subsection is as follows.

Theorem 19 *Let $(\beta t)^{-1} = n\lambda$. Suppose ϵ_j's are i.i.d. random noise with mean zero and finite variance σ_ϵ^2. Let $\tilde{g}_t(\boldsymbol{x}) = \boldsymbol{w}_t^T \mathbf{k}(\boldsymbol{x})$, which is similar to $\hat{f}_t(\boldsymbol{x})$ in (12) but with K_1 instead of K_S and with update rule (11). Then there exists a constant $C > 0$ such that*

$$\mathbb{E}\|\tilde{g}_t - f^*\|_n^2 \leq C\mathbb{E}\|\tilde{g} - f^*\|_n^2, \tag{34}$$

and

$$\mathbb{E}\|\tilde{g}_t\|^2_{\mathcal{H}_{K_1}(\Omega)} \leq 2\mathbb{E}\|\tilde{g}\|^2_{\mathcal{H}_{K_1}(\Omega)}, \tag{35}$$

where the expectation is taken with respect to the noises ϵ_j, $j = 1, ..., n$.

Theorem 19 states that the mean squared prediction error of the early-stopping without weight decay is smaller than (at most the same as) that of the kernel ridge regression estimator, up to a multiplicative constant. This explains why the upper bounds on the early-stopping without weight decay and the kernel ridge regression estimator derived in Raskutti et al. (2014) are identical, in a more explicit way. Note that the conditions of Theorem 19 are quite mild. We do not assume any relationship between f^* and $\mathcal{H}_{K_1}(\Omega)$, and do not require any particular structure of the RKHS $\mathcal{H}_{K_1}(\Omega)$. Furthermore, we do not impose any conditions on λ, and we only require that ϵ_j's are i.i.d. with finite variance (not necessarily sub-Gaussian and can be even heavy-tailed).

It is worth noting that the complexity (i.e., the RKHS norm) of the early-stopping without weight decay is also bounded by the complexity of the kernel ridge regression estimator, up to a constant multiplier. Since the difference between the empirical norm $\|\cdot\|_n$ and the L_2 norm depends on the complexity of the estimator, it can be expected that (34) still holds if we replace the empirical norm by the L_2 norm.

Appendix D. Proof of Theorem 8

In this section, we show the proof of the following theorem. Note that the second statement in Theorem 20 is Theorem 8.

Theorem 20 (Polynomial smoothing) *Suppose Assumptions 1, 2, 4 (C1), and 5 are satisfied. Suppose there exists Ω_1 with positive Lebesgue measure and a Lipschitz boundary such that $\Omega \subset \Omega_1$ and $f^* \in \mathcal{W}^{m_f}(\Omega_1)$. Let $f_{\mathrm{t}}(\boldsymbol{x})$ be as in (12) and $\beta = n^{-1}C_1$ with the positive constant $C_1 \leq 2^{-1}\sup_{\boldsymbol{x}\in\mathbb{R}^D} K_S(\boldsymbol{x})$. Suppose the smoothing scale $\sigma_n \asymp n^{\nu}$ with $\nu \leq 0$. Suppose one of the following holds:*

1. *There is no weight decay in the gradient descent, and the iteration number t satisfies $t \asymp n^{\frac{2(m_0+m_\varepsilon)}{2m_f+d}}\sigma_n^{2m_\varepsilon}$*
2. *There is weight decay in the gradient descent with $\alpha \asymp n^{-1-\frac{2(m_0+m_\varepsilon)}{2m_f+d}}\sigma_n^{-2m_\varepsilon}$, and the iteration number satisfies $t \geq C_2(\frac{m_f}{2m_f+d} + 1/2)\log n/(\log(1-\alpha))$ for some positive constants C_2.*

Then the following statements are true with $N > N_0$, where N is the number of augmentations, and N_0 depends on n and the iteration number t.

1. *For any $a > 0$, there exists an m_ε such that when*
$$\nu = \begin{cases} -\frac{2(2m_0+2m_\varepsilon)D-(2m_0+2m_\varepsilon-D)d}{(2m_f+d)(4m_\varepsilon D-(2m_0+2(1-d^{-1}(2m_f+d)a)m_\varepsilon-D)d)}, & D > d,\\ 0, & D = d,\end{cases}$$
we have
$$\|f_t - f^*\|^2_{L_2(P_{\mathbf{X}})} = O_{\mathbb{P}}\left(n^{-\frac{2m_f}{2m_f+d}+a}\right).$$
2. *Set $m_\varepsilon = 2d^{-1}(2D\max(m_0,m_f)+m_0d)\log n - m_0$. Then by choosing*
$$\nu = \begin{cases} -\frac{2(2m_0+2m_\varepsilon)D-(2m_0+2m_\varepsilon-D)d}{(2m_f+d)(4m_\varepsilon D-(2m_0+2(1-(\log n)^{-1})m_\varepsilon-D)d)} < 0, & D > d,\\ 0, & D = d,\end{cases}$$
we have
$$\|f_t - f^*\|^2_{L_2(P_{\mathbf{X}})} = O_{\mathbb{P}}\left(n^{-\frac{2m_f}{2m_f+d}}(\log n)^{2m_f+1}\right).$$

We first present several lemmas used in this proof. The proof of these lemmas can be found in Appendix I.

Lemma 21 *Suppose the conditions of Theorem 8 are fulfilled. Let f_n^* be the solution to the optimization problem*
$$\min_{g\in\mathcal{H}_{\tilde{K}_S}(\Omega)} \|f^* - g\|^2_{L_2(P_{\mathbf{X}})} + \lambda_n\|g\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)}. \tag{36}$$

Then if $m_0 \leq m_f$, we have

$$\|f^* - f_n^*\|_{L_2(P_{\mathbf{X}})}^2 + \lambda_n \|f_n^*\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)}^2 \leq C_1 \max\left((\lambda_n (m_\varepsilon + 1)^{m_\varepsilon} \sigma_n^{2m_\varepsilon})^{\frac{m_f}{m_0 + m_\varepsilon}}, \lambda_n \right). \tag{37}$$

and if $m_0 > m_f$, we have

$$\|f^* - f_n^*\|_{L_2(P_{\mathbf{X}})}^2 + \lambda_n \|f_n^*\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)}^2 \leq C_2 \max\left((\lambda_n (m_\varepsilon + 1)^{m_\varepsilon} \sigma_n^{2m_\varepsilon})^{\frac{m_f}{m_0 + m_\varepsilon}}, \lambda_n^{\frac{m_f}{m_0}} \right). \tag{38}$$

Here the constants C_1 and C_2 are independent with m_ε.

Lemma 22 *Suppose the conditions of Theorem 8 are fulfilled. Let f_n^* be as in Lemma 21. Suppose there exists $T > 0$ (depending on n) such that*

$$\|f^* - f_n^*\|_{L_2(P_{\mathbf{X}})}^2 + \lambda_n \|f_n^*\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)}^2 \leq T.$$

Let $\hat{f}_n$ be the solution to the optimization problem

$$\min_{g \in \mathcal{H}_{\tilde{K}_S}(\Omega)} \|\boldsymbol{y} - g\|_n^2 + \lambda_n \|g\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)}^2, \tag{39}$$

where $\boldsymbol{y} = (y_1, ..., y_n)^T$. Suppose

$$\sigma_n^{-d/2} m^{\frac{mD}{2m-D} + \frac{1}{2}} \log p$$

converges to zero as n goes to infinity, where $p = \frac{4D}{2m-D}$, and $m = m_0 + m_\varepsilon$. Then we have

$$\begin{aligned}
M_1 = &\max\Bigg((T + n^{-1/2}T^{1/2})^{1/2}, \lambda_n^{-\frac{p}{2(4-p)}} \left(\sigma_n^{-d/2} n^{-1/2} m^{\frac{mD}{2m-D} + \frac{1}{2}} (T + n^{-1/2}T^{1/2})^{\frac{1}{2} - \frac{p}{4}} \right)^{\frac{2}{4-p}}, \\
&\sigma_n^{-d/2} n^{-1/2} m^{\frac{mD}{2m-D} + \frac{1}{2}} \lambda_n^{-\frac{p}{4}}, \left(\sigma_n^{-d/2} n^{-1/2} m^{\frac{mD}{2m-D} + \frac{1}{2}} (\lambda_n^{-1} T)^{\frac{p}{2}} (T + n^{-1/2}T^{1/2})^{1 - \frac{p}{2}} \right)^{1/2}, \\
&(\sigma_n^{-d/2} n^{-1/2} m^{\frac{mD}{2m-D} + \frac{1}{2}})^{\frac{2}{2+p}} (\lambda_n^{-1} T)^{\frac{p}{2(2+p)}} \Bigg), \\
M_2 = &\max\Bigg((\lambda_n^{-1}(T + n^{-1/2}T^{1/2}))^{1/2}, \left(\lambda_n^{-1} \sigma_n^{-d/2} n^{-1/2} m^{\frac{mD}{2m-D} + \frac{1}{2}} (T + n^{-1/2}T^{1/2})^{\frac{1}{2} - \frac{p}{4}} \right)^{\frac{2}{4-p}}, \\
&\sigma_n^{-d/2} n^{-1/2} m^{\frac{mD}{2m-D} + \frac{1}{2}} \lambda_n^{-\frac{2+p}{4}}, \left(\lambda_n^{-1} \sigma_n^{-d/2} n^{-1/2} m^{\frac{mD}{2m-D} + \frac{1}{2}} (\lambda_n^{-1} T)^{\frac{p}{2}} (T + n^{-1/2}T^{1/2})^{1 - \frac{p}{2}} \right)^{1/2}, \\
&\lambda_n^{-1/2} (\sigma_n^{-d/2} n^{-1/2} m^{\frac{mD}{2m-D} + \frac{1}{2}})^{\frac{2}{2+p}} (\lambda_n^{-1} T)^{\frac{p}{2(2+p)}} \Bigg),
\end{aligned}$$

Then we have

$$\|f^* - \hat{f}_n\|_n = O_{\mathbb{P}}(M_1), \|\hat{f}_n\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)} = O_{\mathbb{P}}(M_2).$$

Furthermore, if $\tilde{f}_n$ be the solution to the optimization problem

$$\min_{f \in \mathcal{H}_{\tilde{K}_S}(\Omega)} \|f^* - f\|_n^2 + \lambda_n \|f\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)}, \tag{40}$$

then

$$\|f^* - \tilde{f}_n\|_n = O_{\mathbb{P}}((T + n^{-1/2}T^{1/2})^{1/2}), \|\tilde{f}_n\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)} = O_{\mathbb{P}}((\lambda_n^{-1}(T + n^{-1/2}T^{1/2}))^{1/2}).$$

Lemma 23 (Lemma F.5 of Wang, 2021) *Assume for class $\mathcal{G}$, $\sup_{g\in\mathcal{G}}\|g\|_{L_\infty(\Omega)} \le c < 1$, and the bracket entropy $H_B(\delta_n, \mathcal{G}, \|\cdot\|_{L_2(P_{\mathbf{X}})}) \le \frac{n\delta_n^2}{1200c^2}$, and $n\delta_n^2 \to \infty$, where $0 < \delta_n < 1$. Then we have*

$$P\left(\inf_{\|g\|_{L_2(P_{\mathbf{X}})}\ge 2\delta_n, g\in\mathcal{G}} \frac{\|g\|_n^2}{\|g\|_{L_2(P_{\mathbf{X}})}^2} < C_3\right) \le C_5\exp(-C_6 n\delta_n^2/c^2),$$

and

$$P\left(\sup_{\|g\|_{L_2(P_{\mathbf{X}})}\ge 2\delta_n, g\in\mathcal{G}} \frac{\|g\|_n^2}{\|g\|_{L_2(P_{\mathbf{X}})}^2} > C_4\right) \le C_7\exp(-C_8 n\delta_n^2/c^2),$$

for some constants $C_3, C_4 > 0$ and C_i's ($i = 5,6,7,8$) are only depending on Ω.

Lemma 24 (Interpolation inequality for Polynomial RKHS) *Let $g \in \mathcal{W}^m(\mathbb{R}^D)$. When $r = \frac{D}{2(m_0+m_\varepsilon)}$ and $D > 1$, we have*

$$\|g\|_{L_\infty(\mathbb{R}^D)} \le C_9\|g\|_{L_2(\mathbb{R}^D)}^{1-r}\|g\|_{\mathcal{W}^m(\mathbb{R}^D)}^r,$$

where the positive constant $C_9 = \left(\int_{\mathbb{R}^D}(1+\|\boldsymbol{\omega}\|_2^2)^{-\frac{D}{2}}d\boldsymbol{\omega}\right)^{\frac{1}{2}} < \infty$.

D.1 Without Weight Decay

By the triangle inequality, it can be seen that

$$\|f_t - f^*\|_{L_2(P_{\mathbf{X}})} \le \|f_t - g_t\|_{L_2(P_{\mathbf{X}})} + \|g_t - f^*\|_{L_2(P_{\mathbf{X}})}, \tag{41}$$

where g_t is as in (29).

By Lemma 17, the first term $\|f_t - g_t\|_{L_2(P_{\mathbf{X}})}$ in (41) can be bounded by

$$\|f_t - g_t\|_{L_2(P_{\mathbf{X}})} \le C_{10}\|f_t - g_t\|_{L_\infty(\Omega)} = O_{\mathbb{P}}\left(\frac{n^2\sqrt{\log N/N}}{\eta_n(\tilde{\mathbf{K}})^2}\right),$$

as long as

$$\frac{1}{2}\eta_n(\tilde{\mathbf{K}}) \ge n\sqrt{\frac{\log N}{N}}. \tag{42}$$

Choose

$$N_0 = \frac{4n^2}{\eta_n(\tilde{\mathbf{K}})^2}. \tag{43}$$

Then it holds that when $N \ge N_0$,

$$\|f_t - g_t\|_{L_2(P_{\mathbf{X}})} = O_{\mathbb{P}}\left(n^{-1/2}\right). \tag{44}$$

It remains to consider $\|g_t - f^*\|_{L_2(P_{\mathbf{X}})}$ in (41). In order to do so, we consider the empirical version of $\|g_t - f^*\|_{L_2(P_{\mathbf{X}})}$, and let

$$J_2 = \|g_t - f^*\|_n^2 = \frac{1}{n}\|g_t(\mathbf{X}) - f^*(\mathbf{X})\|_2^2. \tag{45}$$

Let $(\beta t)^{-1} = n\lambda_n$. Consider the kernel ridge regression

$$\tilde{g} = \underset{f\in\mathcal{H}_{\tilde{K}_S}(\Omega)}{\operatorname{argmin}} \|f - \boldsymbol{y}\|_n^2 + \lambda_n\|f\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)}^2. \tag{46}$$

By the representer theorem, $\tilde{g}(\boldsymbol{x}) = \tilde{\mathbf{k}}(\boldsymbol{x})^T(\tilde{\mathbf{K}} + n\lambda_n\mathbf{I})^{-1}\boldsymbol{y}$ for all $\boldsymbol{x}\in\Omega$, where $\tilde{\mathbf{k}}(\boldsymbol{x}) = (\tilde{K}_S(\boldsymbol{x}-\boldsymbol{x}_1),...,\tilde{K}_S(\boldsymbol{x}-\boldsymbol{x}_n))^T$. Then it can be seen that

$$\tilde{g}(\mathbf{X}) - f^*(\mathbf{X}) = n\lambda_n(\tilde{\mathbf{K}} + n\lambda_n\mathbf{I})^{-1}f^*(\mathbf{X}) + \tilde{\mathbf{K}}(\tilde{\mathbf{K}} + n\lambda_n\mathbf{I})^{-1}\boldsymbol{\epsilon} = \boldsymbol{q}_1 + \boldsymbol{q}_2.$$

Recall that (see Equation 29)

$$g_t(\mathbf{X}) = \left(\mathbf{I} - (\mathbf{I} - \beta\tilde{\mathbf{K}})^t\right)\boldsymbol{y},$$

which implies

$$g_t(\mathbf{X}) - f^*(\mathbf{X}) = -(\mathbf{I} - \beta\tilde{\mathbf{K}})^t f^*(\mathbf{X}) + \left(\mathbf{I} - (\mathbf{I} - \beta\tilde{\mathbf{K}})^t\right)\boldsymbol{\epsilon}. \tag{47}$$

By the Cauchy-Schwarz inequality, (45), and (47), it can be seen that

$$\begin{aligned} nJ_2 \le& 2(f^*(\mathbf{X}))^T(\mathbf{I} - \beta\tilde{\mathbf{K}})^{2t}f^*(\mathbf{X}) + 2\boldsymbol{\epsilon}^T(\mathbf{I} - (\mathbf{I} - \beta\tilde{\mathbf{K}})^t)^2\boldsymbol{\epsilon} \\ =& 2nJ_{21} + 2nJ_{22}, \end{aligned} \tag{48}$$

and

$$\begin{aligned} n\|\tilde{g} - f^*\|_n^2 \le& 2(n\lambda_n)^2(f^*(\mathbf{X}))^T(\tilde{\mathbf{K}} + n\lambda_n\mathbf{I})^{-2}f^*(\mathbf{X}) + 2\boldsymbol{\epsilon}^T(\tilde{\mathbf{K}} + n\lambda_n\mathbf{I})^{-1}\tilde{\mathbf{K}}^2(\tilde{\mathbf{K}} + n\lambda_n\mathbf{I})^{-1}\boldsymbol{\epsilon} \\ =& 2\|\boldsymbol{q}_1\|_2^2 + 2\|\boldsymbol{q}_2\|_2^2. \end{aligned} \tag{49}$$

Similar to (90), it can be seen that

$$2nJ_{21} \le C_{11}\|\boldsymbol{q}_1\|_2^2, \tag{50}$$

for some positive constants C_{11}, and similar to (94), the term $2nJ_{22}$ can be further bounded by

$$\begin{aligned} 2nJ_{22} =& 2\sum_{j=1}^n(1 - (1-\beta\eta_j)^t)^2(\boldsymbol{v}_j^T\boldsymbol{\epsilon})^2 \le 2\sum_{j=1}^n\frac{4(\beta t\eta_j)^2}{(1+\beta t\eta_j)^2}(\boldsymbol{v}_j^T\boldsymbol{\epsilon})^2 \\ =& 8\boldsymbol{\epsilon}^T(\tilde{\mathbf{K}} + (\beta t)^{-1}\mathbf{I})^{-1}\tilde{\mathbf{K}}^2(\tilde{\mathbf{K}} + (\beta t)^{-1}\mathbf{I})^{-1}\boldsymbol{\epsilon} = 8\|\boldsymbol{q}_2\|_2^2, \end{aligned} \tag{51}$$

where $\eta_1 \ge \ldots \ge \eta_n > 0$ and $\boldsymbol{v}_j$, $j = 1,\ldots,n$ be the eigenvalues and corresponding eigenvectors of $\tilde{\mathbf{K}}$, respectively. In the last inequality of (51), we note $(\beta t)^{-1} = n\lambda_n$.

Plugging (50) and (51) into (48), we obtain

$$J_2 \leq \frac{2C_{12}}{n} \left(\|\boldsymbol{q}_1\|_2^2 + \|\boldsymbol{q}_2\|_2^2 \right), \tag{52}$$

for some positive constants C_{12}. The term $\|\boldsymbol{q}_1\|_2^2$ and $\|\boldsymbol{q}_2\|_2^2$ can be directly bounded by Lemma 22. To see this, let $f_0(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in \Omega$. Then it can be checked that

$$\frac{1}{n}\|\boldsymbol{q}_1\|_2^2 = \|\tilde{f}_n - f\|_n^2,$$

and

$$\frac{1}{n}\|\boldsymbol{q}_2\|_2^2 = \|\hat{f}_{0,n} - f_0\|_n^2,$$

where $\tilde{f}_n$ is as in (40), and $\hat{f}_{0,n}$ is the solution to the optimization problem

$$\min_{g \in \mathcal{H}_{\tilde{K}_S}(\Omega)} \|\boldsymbol{\epsilon} - g\|_n^2 + \lambda_n \|g\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)}^2.$$

Let $\delta_0 \in (0,1)$ such that $4m_\varepsilon D - (2m_0 + 2(1-\delta_0)m_\varepsilon - D)d > 0$. Take

$$\lambda_n \asymp n^{-\frac{2(m_0+m_\varepsilon)}{2m_f+d}} \sigma_n^{-2m_\varepsilon}, \sigma_n \asymp n^{-\frac{2(2m_0+2m_\varepsilon)D - (2m_0+2m_\varepsilon-D)d}{(2m_f+d)(4m_\varepsilon D - (2m_0+2(1-\delta_0)m_\varepsilon - D)d)}}, n^{-1}(\beta t)^{-1} \asymp \lambda_n, \beta \asymp n^{-1}.$$

Therefore, if $m_\varepsilon = O((\log n)^C)$ for some constant C, and

$$\begin{aligned}
&\lambda_n \leq C_{13}(\lambda_n (m_\varepsilon+1)^{m_\varepsilon} \sigma_n^{2m_\varepsilon})^{\frac{m_f}{m_0+m_\varepsilon}} \\
\Leftrightarrow & n^{-\frac{2(m_0+m_\varepsilon)}{2m_f+d}} n^{\frac{4m_\varepsilon(2m_0+2m_\varepsilon)D - 2m_\varepsilon(2m_0+2m_\varepsilon-D)d}{(2m_f+d)(4m_\varepsilon D-(2m_0+2(1-\delta_0)m_\varepsilon-D)d)}} \leq C_{14} n^{-\frac{2m_f}{2m_f+d}} (m_\varepsilon+1)^{\frac{m_\varepsilon m_f}{m_0+m_\varepsilon}} \\
\Leftarrow & m_\varepsilon^2 \delta_0 d > m_\varepsilon(2m_f D + (m_0 - m_f)(1-\delta_0)d) \\
\Leftarrow & m_\varepsilon > \frac{2m_f D + m_0 d}{\delta_0 d},
\end{aligned} \tag{53}$$

for some positive constants C_{13} and C_{14}, when $m_0 \leq m_f$, or

$$\begin{aligned}
&\lambda_n^{\frac{m_f}{m_0}} \leq C_{15}(\lambda_n (m_\varepsilon+1)^{m_\varepsilon} \sigma_n^{2m_\varepsilon})^{\frac{m_f}{m_0+m_\varepsilon}} \\
\Leftrightarrow & n^{-\frac{2(m_0+m_\varepsilon)}{2m_f+d}} n^{\frac{4m_\varepsilon(2m_0+2m_\varepsilon)D - 2m_\varepsilon(2m_0+2m_\varepsilon-D)d}{(2m_f+d)(4m_\varepsilon D-(2m_0+2(1-\delta_0)m_\varepsilon-D)d)}} \leq C_{16} n^{-\frac{2m_0}{2m_f+d}} (m_\varepsilon+1)^{\frac{m_\varepsilon m_0}{m_0+m_\varepsilon}} \\
\Leftarrow & m_\varepsilon^2 \delta_0 d > 2m_0 m_\varepsilon D \\
\Leftarrow & m_\varepsilon > \frac{2m_0 D + m_0 d}{\delta_0 d},
\end{aligned} \tag{54}$$

for some positive constants C_{15} and C_{16}, when $m_0 > m_f$, we have

$$T \leq C_{17} n^{-\frac{2m_f}{2m_f+d}} (m_\varepsilon+1)^{\frac{m_\varepsilon m_f}{m_0+m_\varepsilon}} \leq C_{17} n^{-\frac{2m_f}{2m_f+d}} (m_\varepsilon+1)^{m_f},$$

for some positive constants C_{17}, where T is as in Lemma 22. Suppose $D > 1$, long but tedious calculation shows that

$$M_1 \leq C_{18}(m_\varepsilon + m_0)^{m_f+\frac{1}{2}} n^{-\frac{m_f}{2m_f+d}+\delta'},$$

for some positive constants C_{18}, where M_1 is as in Lemma 22, and

$$\delta' = \frac{((4m_0 + 4m_\varepsilon)D - (2m_0 + 2m_\varepsilon - D)d)m_\varepsilon d}{(2m_f + d)(2m_\varepsilon + 2m_0 - D)(4m_\varepsilon D - (2m_0 + 2(1-\delta_0)m_\varepsilon - D)d)}\delta_0 \leq \frac{d}{2(2m_f + d)}\delta_0,$$

where the inequality is because of (53) (if $m_0 \leq m_f$) or (54) (if $m_0 > m_f$). Therefore, by taking $\delta_0 = d^{-1}(2m_f + d)a$ and $m_\varepsilon = (\delta_0 d)^{-1}(2D\max(m_0, m_f) + m_0 d) + 1$, we have

$$\begin{aligned}\frac{1}{n}\|\boldsymbol{q}_1\|_2^2 =& \|\tilde{f}_n - f\|_n^2 = O_{\mathbb{P}}\left(n^{-\frac{2m_f}{2m_f+d}+a}\right),\\ \frac{1}{n}\|\boldsymbol{q}_2\|_2^2 =& \|\hat{f}_{0,n} - f_0\|_n^2 = O_{\mathbb{P}}\left(n^{-\frac{2m_f}{2m_f+d}+a}\right).\end{aligned} \tag{55}$$

Then by (52) and (55), we obtain

$$J_2 = O_{\mathbb{P}}\left(n^{-\frac{2m_f}{2m_f+d}+a}\right), \tag{56}$$

which corresponds to the first statement of Theorem 8.

Taking $\delta_0 = (\log n)^{-1}$, we obtain that

$$M_1 \leq C_{18} n^{-\frac{m_f}{2m_f+d}} e^{\frac{d}{2(2m_f+d)}}(m_\varepsilon + m_0)^{m_f+\frac{1}{2}} \leq C_{19} n^{-\frac{m_f}{2m_f+d}}(m_\varepsilon + m_0)^{m_f+\frac{1}{2}},$$

for some positive constants C_{19}, where we require $m_\varepsilon > d^{-1}(2D\max(m_0, m_f) + m_0 d)\log n$. Thus, we can directly take $m_\varepsilon = 2d^{-1}(2D\max(m_0, m_f) + m_0 d)\log n - m_0$ such that

$$\begin{aligned}\frac{1}{n}\|\boldsymbol{q}_1\|_2^2 =& \|\tilde{f}_n - f\|_n^2 = O_{\mathbb{P}}\left(n^{-\frac{2m_f}{2m_f+d}}(\log n)^{2m_f+1}\right),\\ \frac{1}{n}\|\boldsymbol{q}_2\|_2^2 =& \|\hat{f}_{0,n} - f_0\|_n^2 = O_{\mathbb{P}}\left(n^{-\frac{2m_f}{2m_f+d}}(\log n)^{2m_f+1}\right).\end{aligned} \tag{57}$$

Thus, by (52) and (57), we have

$$J_2 = O_{\mathbb{P}}\left(n^{-\frac{2m_f}{2m_f+d}}(\log n)^{2m_f+1}\right), \tag{58}$$

which corresponds to the second statement of Theorem 8.

It remains to bound $\|g_t - f^*\|_{L_2(P_{\mathbf{X}})}$. Note that

$$\|g_t - f^*\|_{L_2(P_{\mathbf{X}})} \leq \|g_t - f_n^*\|_{L_2(P_{\mathbf{X}})} + \|f_n - f^*\|_{L_2(P_{\mathbf{X}})} \leq \|g_t - f_n^*\|_{L_2(P_{\mathbf{X}})} + T^{1/2},$$

and

$$\begin{aligned}\|g_t - f_n^*\|_n \leq & \|g_t - f^*\|_n + \|f_n^* - f^*\|_n \leq \|g_t - f^*\|_n + O_{\mathbb{P}}\left(\left(T + n^{-1/2}T^{1/2}\right)^{1/2}\right) \\ \leq & O_{\mathbb{P}}\left(n^{-\frac{2m_f}{2m_f+d}}(\log n)^{2m_f+1}\right),\end{aligned}$$

where the second inequality is because of (135). Therefore, it suffices to bound the difference between $\|g_t - f_n^*\|_{L_2(P_{\mathbf{X}})}$ and $\|g_t - f_n^*\|_n$. By (95) and Lemma 22, we have

$$\|g_t\|^2_{\mathcal{N}_\sigma(\Omega)} \leq \sigma_n^{-2m_0}\|g_t\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)} \leq C_{20}\sigma_n^{-2m_0}\|\tilde{g}\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)} = O_{\mathbb{P}}\left(n^{\nu_1}(\log n)^{2m_f+1}\right), \tag{59}$$

for some positive constants C_{20}, where

$$\begin{aligned}\nu_1 = & \frac{2(m_0 + m_\varepsilon - m_f)}{2m_f + d} + 2(m_\varepsilon - m_0)\nu, \\ \text{and } \nu = & -\frac{2(2m_0 + 2m_\varepsilon)D - (2m_0 + 2m_\varepsilon - D)d}{(2m_f + d)(4m_\varepsilon D - (2m_0 + 2(1-\delta_0)m_\varepsilon - D)d)}.\end{aligned} \tag{60}$$

Consider function class $\mathcal{G} = \{h : h = (g_t - f_n^*)/(C_{21}n^{\nu_1/2}(\log n)^{m_f+1/2})\}$, where the constant C_{21} is taken such that $\|h_1\|_{\mathcal{N}_\sigma(\Omega)} < 1$ for all $h_1 \in \mathcal{G}$. Then lemma 24 leads to

$$\|h_1\|_{L_\infty(\Omega)} \leq C_{22}\|h_1\|_{L_2(P_{\mathbf{X}})}^{1-\frac{D}{2(m_0+m_\varepsilon)}}\|h_1\|_{\mathcal{N}_\sigma(\Omega)}^{\frac{D}{2(m_0+m_\varepsilon)}},$$

for some positive constants C_{22} and all $h_1 \in \mathcal{G}$, which implies

$$c_1 := \sup_{h_1\in\mathcal{G}}\|h_1\|_{L_\infty(\Omega)} \leq C_{22}R_1^{1-\frac{D}{2(m_0+m_\varepsilon)}},$$

where $R_1 = \sup_{h_1\in\mathcal{G}}\|h_1\|_{L_2(P_{\mathbf{X}})} \leq \sup_{h_1\in\mathcal{G}}\|h_1\|_{L_\infty(\Omega)} \leq \sup_{h_1\in\mathcal{G}}\|h_1\|_{\mathcal{N}_\sigma(\Omega)} < 1$, because of the reproducing property. Let $m = m_0 + m_\varepsilon$. Taking $c = C_{22}R_1^{1-\frac{D}{2m}} < 1$, and $\delta_n = C_{23}(\sigma_n^{-d}n^{-1}c^2m^{\frac{2mD}{2m-D}})^{\frac{2m-D}{4m}}$ for some positive constants C_{23} in Lemma 23, it can be checked that

$$C_{24}n\delta_n^2c^{-2} \geq H(\delta, \mathcal{B}_{\mathcal{H}_\sigma(\Omega)}, \|\cdot\|_{L_\infty(\Omega)}),$$

for some positive constants C_{24}, which implies the conditions of Lemma 23 are fulfilled. Applying Lemma 23 to the case $\|g_t - f_n^*\|^2_{L_2(P_{\mathbf{X}})} \geq \delta_n^2n^{\nu_1}$, together with (58), we have

$$R_1 = O_{\mathbb{P}}\left(\max\{n^{-\frac{m_f}{2m_f+d}-\nu_1/2}(\log n)^{m_f+1/2}, \delta_n\}\right). \tag{61}$$

If $\delta_n \geq n^{-\frac{m_f}{2m_f+d}-\nu_1/2}(\log n)^{m_f+1/2}$, we have $R_1 \leq C_{25}\delta_n$ for some positive constants C_{25}, which implies

$$R_1 \leq C_{26}(\sigma_n^{-d}n^{-1}c^2m^{\frac{2mD}{2m-D}})^{\frac{2m-D}{4m}},$$

for some positive constants C_{26}. Therefore, we have

$$\|g_t - f_n^*\|_{L_2(P_{\mathbf{X}})} \leq C_{21} n^{\nu_1/2} R_1 \leq C_{27} n^{\nu_2} (\log n)^{D/2},$$

where

$$\nu_2 = \frac{(m_0 + m_\varepsilon - m_f)}{2m_f + d} + (m_\varepsilon - m_0)\nu - \frac{2m - D}{4m}(d\nu + 1) < -\frac{m_f}{2m_f + d}.$$

If $\delta_n < n^{-\frac{m_f}{2m_f+d} - \nu_1/2} (\log n)^{m_f + 1/2}$, then $R_1 = O_{\mathbb{P}}(n^{-\frac{m_f}{2m_f+d} - \nu_1/2} (\log n)^{m_f+1/2})$, which implies $\|g_t - f_n^*\|_{L_2(P_{\mathbf{X}})} = O_{\mathbb{P}}(n^{-\frac{m_f}{2m_f+d}} (\log n)^{m_f+1/2})$. Here we note that the proof is still valid if we replace g_t with $\tilde{g}$. Therefore, in both cases we have $\|g_t - f_n^*\|_{L_2(P_{\mathbf{X}})} = O_{\mathbb{P}}(n^{-\frac{m_f}{2m_f+d}} (\log n)^{m_f+1/2})$, which, together with (51) and (44), finishes the proof. ■

D.2 With Weight Decay

If $\alpha > 0$, we decompose the error by

$$\begin{aligned}\|f_t - f^*\|_{L_2(P_{\mathbf{X}})} \leq & \|f_t - g_t\|_{L_2(P_{\mathbf{X}})} + \|\tilde{\mathbf{k}}(\cdot)^T (\alpha/\beta \mathbf{I} + \tilde{\mathbf{K}})^{-1} \boldsymbol{y} - f^*\|_{L_2(P_{\mathbf{X}})} \\ & + \|\beta \mathbf{k}(\cdot)^T ((1-\alpha)\mathbf{I} - \beta \tilde{\mathbf{K}})^t (\alpha \mathbf{I} + \beta \tilde{\mathbf{K}})^{-1} \boldsymbol{y}\|_{L_2(P_{\mathbf{X}})} \\ = & I_1 + I_2 + I_3. \end{aligned} \tag{62}$$

As in (44), there exists an N_0 (depending on n) such that when $N \geq N_0$,

$$I_1 = O_{\mathbb{P}}\left(n^{-1/2}\right). \tag{63}$$

The second term is the error $\|\tilde{f}_n - f^*\|_{L_2(P_{\mathbf{X}})}$, where $\tilde{f}_n$ is as in (40). Lemma 22 gives us that

$$\|\tilde{f}_n - f^*\|_n = O_{\mathbb{P}}(n^{-\frac{m_f}{2m_f+d}}).$$

Following a similar approach in Appendix D.1, it can be further shown that

$$I_2 = O_{\mathbb{P}}(n^{-\frac{m_f}{2m_f+d}}), \tag{64}$$

where we let $\alpha \asymp n^{-1-\frac{2(m_0+m_\varepsilon)}{2m_f+d}} \sigma_n^{-2m_\varepsilon}$, and β and σ_n are as in Theorem 8.

It remains to bound I_3 in (62). By Cauchy-Schwarz inequality,

$$\begin{aligned}
&\|\beta\mathbf{k}(\cdot)^T((1-\alpha)\mathbf{I}-\beta\tilde{\mathbf{K}})^t(\alpha\mathbf{I}+\beta\tilde{\mathbf{K}})^{-1}\boldsymbol{y}\|_{L_2(P_{\mathbf{X}})}\\
\leq&\left\|\left(\operatorname{tr}\left(\left((\alpha/\beta\mathbf{I}+\tilde{\mathbf{K}})^{-1}\boldsymbol{y}\mathbf{k}(\cdot)^T\right)^2\right)\operatorname{tr}\left(((1-\alpha)\mathbf{I}-\beta\tilde{\mathbf{K}})^{2t}\right)\right)^{1/2}\right\|_{L_2(P_{\mathbf{X}})}\\
\leq&\left\|\mathbf{k}(\cdot)^T(\alpha/\beta\mathbf{I}+\tilde{\mathbf{K}})^{-1}\boldsymbol{y}\right\|_{L_2(P_{\mathbf{X}})}\left(\operatorname{tr}\left(((1-\alpha)\mathbf{I}-\beta\tilde{\mathbf{K}})^{2t})\right)\right)^{1/2}\\
\leq&\|(\mathbf{k}(\cdot)^T\mathbf{k}(\cdot))^{1/2}\|_{L_2(P_{\mathbf{X}})}\|\boldsymbol{y}\|_2\beta/\alpha\\
=&O_{\mathbb{P}}\left(n^{1+\frac{2(m_0+m_\varepsilon)}{2m_f+d}}\sigma_n^{2m_\varepsilon}(1-\alpha)^t\right).
\end{aligned}\tag{65}$$

Thus, there exists $t_0>0$ such that as long as $t>t_0$, I_2 dominates I_3. Combining (63), (64), and (65), we finish the proof. ■

Appendix E. Proof of Theorem 9

We first present some lemmas, whose proofs can be found in Appendix J.

Lemma 25 *Let $k_\sigma(\boldsymbol{x}-\boldsymbol{x}')$ be a Gaussian kernel defined by*

$$k_\sigma(\boldsymbol{x}-\boldsymbol{x}')=\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|_2^2}{4\sigma^2}\right),\tag{66}$$

and $\mathcal{H}_\sigma(\mathbb{R}^D)$ be the RKHS generated by $k_\sigma(\boldsymbol{x}-\boldsymbol{x}')$. Then we have

$$\|h_1\|_{\mathcal{H}_{\sigma_n/\sqrt{2}}(\mathbb{R}^D)}\leq C_1\sigma_n^{-D/2}\|h_1\|_{\mathcal{H}_{\tilde{K}_S}(\mathbb{R}^D)},$$

and

$$\|h_2\|_{\mathcal{H}_{\tilde{K}_S}(\mathbb{R}^D)}\leq C_2\sigma_n^{-m_0-D/2}\|h_2\|_{\mathcal{H}_{\sqrt{3}\sigma_n}(\mathbb{R}^D)},$$

for $h_1\in\mathcal{H}_{\tilde{K}_S}(\mathbb{R}^D)$ and $h_2\in\mathcal{H}_{\sqrt{3}\sigma_n}(\mathbb{R}^D)$, where the positive constants C_1 and C_2 does not depend on σ_n.

Lemma 26 *Let f_n^* be the solution to the optimization problem*

$$\min_{g\in\mathcal{H}_{\tilde{K}_S}(\Omega)}\|f^*-g\|^2_{L_2(P_{\mathbf{X}})}+\lambda_n\|g\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)}.\tag{67}$$

Then

$$\|f^*-f_n^*\|^2_{L_2(P_{\mathbf{X}})}\leq C_3\max(\lambda_n\sigma_n^{-2m_0},\sigma_n^{2m_f}),$$

and

$$\|f_n^*\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)}\leq C_3\lambda_n^{-1}\max(\lambda_n\sigma_n^{-2m_0},\sigma_n^{2m_f}),$$

for some positive constants C_3.

Lemma 27 *Let f_n^* be the solution to the optimization problem*

$$\min_{g\in\mathcal{H}_{\tilde{K}_S}(\Omega)} \|f^*-g\|_{L_2(P_{\mathbf{X}})}^2+\lambda_n\|g\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)}^2. \tag{68}$$

Suppose there exists $T>0$ (depending on n) such that

$$\|f^*-f_n^*\|_{L_2(P_{\mathbf{X}})}^2+\lambda_n\|f_n^*\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)}^2\leq T.$$

Let $\hat{f}_n$ be the solution to the optimization problem

$$\min_{g\in\mathcal{H}_{\tilde{K}_S}(\Omega)} \|\boldsymbol{y}-g\|_n^2+\lambda_n\|g\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)}^2. \tag{69}$$

Let $p=(\log n)^{-1}$,

$$\begin{aligned}
M_1 =\max\Bigg(&(T+n^{-1/2}T^{1/2})^{1/2},\sigma_n^{-d/2-\frac{pD}{4}}p^{-(D+1)/2}n^{-1/2}\lambda_n^{-\frac{p}{4}},\\
&\lambda_n^{-\frac{p}{2(4-p)}}\left(\sigma_n^{-d/2-\frac{pD}{4}}p^{-(D+1)/2}n^{-1/2}(T+n^{-1/2}T^{1/2})^{\frac{1}{2}-\frac{p}{4}}\right)^{\frac{2}{4-p}},\\
&\left(\sigma_n^{-d/2-\frac{pD}{4}}p^{-(D+1)/2}n^{-1/2}(\lambda_n^{-1}T)^{\frac{p}{2}}(T+n^{-1/2}T^{1/2})^{1-\frac{p}{2}}\right)^{1/2},\\
&(\sigma_n^{-d/2-\frac{pD}{4}}p^{-(D+1)/2}n^{-1/2})^{\frac{2}{2+p}}(\lambda_n^{-1}T)^{\frac{p}{2+p}}\Bigg),\\
M_2 =\max\Bigg(&(\lambda_n^{-1}(T+n^{-1/2}T^{1/2}))^{1/2},\sigma_n^{-d/2-\frac{pD}{4}}p^{-(D+1)/2}n^{-1/2}\lambda_n^{-\frac{2+p}{4}},\\
&\left(\lambda_n^{-1}\sigma_n^{-d/2-\frac{pD}{4}}p^{-(D+1)/2}n^{-1/2}(T+n^{-1/2}T^{1/2})^{\frac{1}{2}-\frac{p}{4}}\right)^{\frac{2}{4-p}},\\
&\left(\lambda_n^{-1}\sigma_n^{-d/2-\frac{pD}{4}}p^{-(D+1)/2}n^{-1/2}(\lambda_n^{-1}T)^{\frac{p}{2}}(T+n^{-1/2}T^{1/2})^{1-\frac{p}{2}}\right)^{1/2},\\
&\lambda_n^{-1/2}(\sigma_n^{-d/2-\frac{pD}{4}}p^{-(D+1)/2}n^{-1/2})^{\frac{2}{2+p}}(\lambda_n^{-1}T)^{\frac{p}{2+p}}\Bigg).
\end{aligned}$$

Then we have

$$\|f^*-\hat{f}_n\|_n=O_{\mathbb{P}}(M_1),\|\hat{f}_n\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)}=O_{\mathbb{P}}(M_2).$$

Furthermore, if $\tilde{f}_n$ be the solution to the optimization problem

$$\min_{f\in\mathcal{H}_{\tilde{K}_S}(\Omega)} \|f^*-f\|_n^2+\lambda_n\|f\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)}, \tag{70}$$

then

$$\|f^*-\tilde{f}_n\|_n=O_{\mathbb{P}}((T+n^{-1/2}T^{1/2})^{1/2}),\|\tilde{f}_n\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)}=O_{\mathbb{P}}((\lambda_n^{-1}(T+n^{-1/2}T^{1/2}))^{1/2}).$$

Lemma 28 (Interpolation inequality for Gaussian RKHS) *Let $g \in \mathcal{H}_\sigma(\mathbb{R}^D)$. For any $1 > r > 0$, we have*

$$\|g\|_{L_\infty(\mathbb{R}^D)} \leq C_4 r^{-\frac{D}{4}} \sigma^{\frac{D(r-1)}{2}} \|g\|^{1-r}_{L_2(\mathbb{R}^D)} \|g\|^r_{\mathcal{H}_\sigma(\mathbb{R}^D)},$$

where C_4 is a constant not related to r, σ and g.

E.1 Without Weight Decay

We first decompose the error as

$$\|f_t - f^*\|_{L_2(P_{\mathbf{X}})} \leq \|f_t - g_t\|_{L_2(P_{\mathbf{X}})} + \|g_t - f^*\|_{L_2(P_{\mathbf{X}})}, \tag{71}$$

where g_t is as in (29).

By Lemma 17, the first term $\|f_t - g_t\|_{L_2(P_{\mathbf{X}})}$ in (71) can be bounded by

$$\|f_t - g_t\|_{L_2(P_{\mathbf{X}})} \leq C_5 \|f_t - g_t\|_{L_\infty(\Omega)} = O_{\mathbb{P}}\left(\frac{n^2\sqrt{\log N/N}}{\eta_n(\tilde{\mathbf{K}})^2}\right),$$

for some positive constants C_5, as long as

$$\frac{1}{2}\eta_n(\tilde{\mathbf{K}}) \geq n\sqrt{\frac{\log N}{N}}. \tag{72}$$

Choose

$$N_0 = \frac{4n^2}{\eta_n(\tilde{\mathbf{K}})^2}. \tag{73}$$

Then it holds that when $N \geq N_0$,

$$\|f_t - g_t\|_{L_2(P_{\mathbf{X}})} = O_{\mathbb{P}}\left(n^{-1/2}\right). \tag{74}$$

It remains to consider $\|g_t - f^*\|_{L_2(P_{\mathbf{X}})}$. We consider the empirical version of $\|g_t - f^*\|_{L_2(P_{\mathbf{X}})}$, and let

$$J_2 = \|g_t - f^*\|_n^2 = \frac{1}{n}\|g_t(\mathbf{X}) - f^*(\mathbf{X})\|_2^2. \tag{75}$$

Let $(\beta t)^{-1} = n\lambda_n$. Consider the kernel ridge regression

$$\tilde{g} = \operatorname*{argmin}_{f \in \mathcal{H}_{\tilde{K}_S}(\Omega)} \|f - \boldsymbol{y}\|_n^2 + \lambda_n \|f\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)}.$$

By the representer theorem, $\tilde{g}(\boldsymbol{x}) = \tilde{\mathbf{k}}(\boldsymbol{x})^T(\tilde{\mathbf{K}} + n\lambda_n\mathbf{I})^{-1}\boldsymbol{y}$ for all $\boldsymbol{x} \in \Omega$. Then it can be seen that

$$\tilde{g}(\mathbf{X}) - f^*(\mathbf{X}) = n\lambda_n(\tilde{\mathbf{K}} + n\lambda_n\mathbf{I})^{-1} f^*(\mathbf{X}) + \tilde{\mathbf{K}}(\tilde{\mathbf{K}} + n\lambda_n\mathbf{I})^{-1}\boldsymbol{\epsilon} = \boldsymbol{q}_1 + \boldsymbol{q}_2,$$

Following the arguments in Appendix D.1, the term J_2 can be bounded by

$$J_2 \leq \frac{2}{n}\left(2C_6\|\boldsymbol{q}_1\|_2^2 + 8\|\boldsymbol{q}_2\|_2^2\right) \tag{76}$$

for some positive constants C_6, and

$$\frac{1}{n}\|\boldsymbol{q}_1\|_2^2 = \|\tilde{f}_n - f^*\|_n^2,$$

and

$$\frac{1}{n}\|\boldsymbol{q}_2\|_2^2 = \|\hat{f}_{0,n} - f_0\|_n^2,$$

where $f_0(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in \Omega$, $\tilde{f}_n$ is as in (70), and $\hat{f}_{0,n}$ is the solution to the optimization problem

$$\min_{g \in \mathcal{H}_{\tilde{K}_S}(\Omega)} \|\boldsymbol{\epsilon} - g\|_n^2 + \lambda_n \|g\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)}.$$

By setting $\beta t \asymp n^{\frac{2m_0-d}{2m_f+d}}$ (which implies $\lambda_n \asymp n^{-\frac{2m_0+2m_f}{2m_f+d}}$), $\sigma_n \asymp n^{-\frac{1}{2m_f+d}}$, Lemma 26 implies that $T \asymp n^{-\frac{2m_f}{2m_f+d}}$, which, together with Lemma 27, implies

$$\begin{aligned}
\frac{1}{n}\|\boldsymbol{q}_1\|_2^2 &= \|\tilde{f}_n - f^*\|_n^2 = O_{\mathbb{P}}\left(n^{-\frac{2m_f}{2m_f+d}}(\log n)^{D+1}\right), \\
\frac{1}{n}\|\boldsymbol{q}_2\|_2^2 &= \|\hat{f}_{0,n} - f_0\|_n^2 = O_{\mathbb{P}}\left(n^{-\frac{2m_f}{2m_f+d}}(\log n)^{D+1}\right).
\end{aligned} \tag{77}$$

By (77) and (76), we obtain

$$J_2 = O_{\mathbb{P}}\left(n^{-\frac{2m_f}{2m_f+d}}(\log n)^{D+1}\right). \tag{78}$$

Next, we consider bounding $\|g_t - f^*\|_{L_2(P_{\mathbf{X}})}$. Similar to the proof in Appendix D.1, it suffices to consider bounding the difference between $\|g_t - f_n^*\|_{L_2(P_{\mathbf{X}})}$ and $\|g_t - f_n^*\|_n$. Lemma 25 implies that

$$\|\tilde{g}\|^2_{\mathcal{H}_{\sigma_n/\sqrt{2}}(\Omega)} \leq C_7 \sigma_n^{-D} \|\tilde{g}\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)} = O_{\mathbb{P}}\left(n^{\frac{2m_0+D}{2m_f+d}}(\log n)^{D+1}\right), \tag{79}$$

for some positive constants C_7.

Consider function class $\mathcal{G} = \{h : h = (g_t - f_n^*)/(2C_8 n^{\frac{m_0+D/2}{2m_f+D}}(\log n)^{(D+1)/2})\}$, where the constant C_8 is taken such that $\|h_1\|_{\mathcal{H}_{\sigma_n/\sqrt{2}}(\Omega)} < 1$ for all $h_1 \in \mathcal{G}$. Taking $r = (\log n)^{-1}$ in Lemma 28, together with the extension theorem leads to

$$\|h_1\|_{L_\infty(\Omega)} \leq C_9 r^{-\frac{D}{4}} \sigma_n^{\frac{D(r-1)}{2}} \|h_1\|^{1-r}_{L_2(P_{\mathbf{X}})} \|h_1\|^r_{\mathcal{H}_{\sigma_n/\sqrt{2}}(\Omega)},$$

for some positive constants C_9 and all $h_1 \in \mathcal{G}$. Therefore, we have

$$c_1 := \sup_{h_1 \in \mathcal{G}} \|h_1\|_{L_\infty(\Omega)} \leq C_9 r^{-\frac{D}{4}} \sigma_n^{\frac{D(r-1)}{2}} R_1^{1-r},$$

where $R_1 = \sup_{h_1 \in \mathcal{G}} \|h_1\|_{L_2(P_{\mathbf{X}})} \leq \sup_{h_1 \in \mathcal{G}} \|h_1\|_{L_\infty(\Omega)} \leq \sup_{h_1 \in \mathcal{G}} \|h_1\|_{\mathcal{H}_{\sigma_n/\sqrt{2}}(\Omega)} < 1$, because of the reproducing property. Taking $c = C_9 r^{-\frac{D}{4}} \sigma_n^{\frac{D(r-1)}{2}} R_1^{1-r}$ and $\delta_n = C_{10}(\sigma_n^d r^{-D-1} c^{-2})^{\frac{1}{r+2}}$ for some positive constants C_{10} in Lemma 23, it can be checked that

$$C_{11} n \delta_n^2 c^{-2} \geq H(\delta_n, \mathcal{B}_{\mathcal{H}_{\sigma_n/\sqrt{2}}(\Omega)}, \|\cdot\|_{L_\infty(\Omega)}),$$

for some positive constants C_{11}. By repeating the proof in Appendix D.1, we obtain that

$$\|g_t - f^*\|_{L_2(P_{\mathbf{X}})} = O_{\mathbb{P}}\left(n^{-\frac{2m_f}{2m_f+D}} (\log n)^{D+1}\right),$$

which, together with (71) and (74), implies

$$\|f_t - f^*\|_{L_2(P_{\mathbf{X}})} = O_{\mathbb{P}}\left(n^{-\frac{2m_f}{2m_f+D}} (\log n)^{D+1}\right).$$

This finishes the proof. ■

E.2 With Weight Decay

The results can be obtained by merely repeating the proof in Appendix D.2, where the only difference is that the corresponding convergence rate for I_2 (in Equation 62 of Appendix D.2) is obtained via the proof in Appendix E.1. Thus we omit it here.

Appendix F. Proof of Theorem 13

We first present several lemmas used in this proof.

Lemma 29 *Suppose the conditions of Theorem 13 are fulfilled and $f^* \in \mathcal{MW}^{m_f}(\Omega_1)$. Let f_n^* be the solution to the optimization problem*

$$\min_{g \in \mathcal{H}_{\tilde{K}_S}(\Omega)} \|f^* - g\|_{L_2(P_{\mathbf{X}})}^2 + \lambda_n \|g\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)}^2.$$

Then

$$\|f^* - f_n^*\|_{L_2(P_{\mathbf{X}})}^2 + \lambda_n \|f_n^*\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)}^2 \lesssim \sum_{\boldsymbol{l} \in \{0,1\}^D : |\boldsymbol{l}| \geq 1} (\lambda_n \sigma_n^{2m_\varepsilon |\boldsymbol{l}|})^{\frac{m_f}{m_0 + m_\varepsilon}}.$$

Lemma 30 *Suppose the conditions of Theorem 13 are fulfilled. Let f_n^* be as in Lemma 29. Suppose there exists $T > 0$ (depending on n) such that*

$$\|f^* - f_n^*\|_{L_2(P_{\mathbf{X}})}^2 + \lambda_n \|f_n^*\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)}^2 \leq T.$$

Let $\hat{f}_n$ be the solution to the optimization problem

$$\|\boldsymbol{y} - \hat{f}_n\|_n^2 + \lambda_n \|\hat{f}_n\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)}^2. \tag{80}$$

Let $p = \frac{1}{m_0 + m_\varepsilon}$, $q = \frac{D-1}{2} + \frac{p}{4}$

$$\begin{aligned}
M_1 = \max \Bigg(& \lambda_n^{-\frac{p}{2(4-p)}} \left(\sigma_n^{-d/2} n^{-1/2} (T + n^{-1/2} T^{1/2})^{\frac{1}{2} - \frac{p}{4}} \big| \log(T + n^{-1/2} T^{1/2}) \big|^q \right)^{\frac{2}{4-p}}, \\
& (T + n^{-1/2} T^{1/2})^{1/2}, \sigma_n^{-d/2} n^{-1/2} \lambda_n^{-\frac{p}{4}} \big| \log(\sigma_n^{-d/2} n^{-1/2} \lambda_n^{-\frac{p}{4}}) \big|^q, \\
& \left(\sigma_n^{-d/2} n^{-1/2} (\lambda_n^{-1} T)^{\frac{p}{2}} (T + n^{-1/2} T^{1/2})^{1 - \frac{p}{2}} \big| \log(T + n^{-1/2} T^{1/2}) \big|^q \right)^{1/2}, \\
& (\sigma_n^{-d/2} n^{-1/2})^{\frac{2}{2+p}} (\lambda_n^{-1} T)^{\frac{p}{2(2+p)}} \big| \log \left((\sigma_n^{-d/2} n^{-1/2})^{\frac{2}{2+p}} (\lambda_n^{-1} T)^{\frac{p}{2(2+p)}} \right) \big|^{q \frac{2}{2+p}} \Bigg), \\
M_2 = \max \Bigg(& \left(\lambda_n^{-1} \sigma_n^{-d/2} n^{-1/2} (T + n^{-1/2} T^{1/2})^{\frac{1}{2} - \frac{p}{4}} \big| \log(T + n^{-1/2} T^{1/2}) \big|^q \right)^{\frac{2}{4-p}}, \\
& (\lambda_n^{-1} (T + n^{-1/2} T^{1/2}))^{1/2}, \sigma_n^{-d/2} n^{-1/2} \lambda_n^{-\frac{2+p}{4}} \big| \log(\sigma_n^{-d/2} n^{-1/2} \lambda_n^{-\frac{p}{4}}) \big|^q, \\
& \left(\lambda_n^{-1} \sigma_n^{-d/2} n^{-1/2} (\lambda_n^{-1} T)^{\frac{p}{2}} (T + n^{-1/2} T^{1/2})^{1 - \frac{p}{2}} \big| \log(T + n^{-1/2} T^{1/2})) \big|^q \right)^{1/2}, \\
& \lambda_n^{-1/2} (\sigma_n^{-d/2} n^{-1/2})^{\frac{2}{2+p}} (\lambda_n^{-1} T)^{\frac{p}{2(2+p)}} \left| \log \left((\sigma_n^{-d/2} n^{-1/2})^{\frac{2}{2+p}} (\lambda_n^{-1} T)^{\frac{p}{2(2+p)}} \right) \right|^{\frac{2q}{2+p}} \Bigg).
\end{aligned}$$

Then we have

$$\|f^* - \hat{f}_n\|_n = O_{\mathbb{P}}(M_1), \|\hat{f}_n\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)} = O_{\mathbb{P}}(M_2).$$

Furthermore, if $\tilde{f}_n$ is the solution to the optimization problem

$$\|f^* - \tilde{f}_n\|_n^2 + \lambda_n \|\tilde{f}_n\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)}, \tag{81}$$

then

$$\|f^* - \tilde{f}_n\|_n = O_{\mathbb{P}}((T + n^{-1/2} T^{1/2})^{1/2}), \|\tilde{f}_n\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)} = O_{\mathbb{P}}((\lambda_n^{-1}(T + n^{-1/2} T^{1/2}))^{1/2}).$$

Lemma 31 (Interpolation inequality for tensored RKHS) *Let $g \in \mathcal{MW}^m(\mathbb{R}^D)$. For any $1 \geq r > m^{-1}/2$, we have*

$$\|g\|_{L_\infty(\mathbb{R}^D)} \leq C_r \|g\|_{L_2(\mathbb{R}^D)}^{1-r} \|g\|_{\mathcal{MW}^m(\mathbb{R}^D)}^r,$$

where C_r is a constant that only depends on r.

F.1 Without Weight Decay

The result can be obtained by merely repeating the proof in Appendix D.1. We let $\lambda_n \asymp n^{-\frac{2(m_0+m_\varepsilon)}{2m_f+1}}(\log n)^{\frac{2(D-1)(m_0+m_\varepsilon)+1}{2m_f+1}}$, $\sigma_n \asymp 1$, then by Lemma 29 and Lemma 30, the term J_2 in (58) becomes

$$J_2 = O_{\mathbb{P}}\left(n^{-\frac{2m_f}{2m_f+1}}(\log n)^{\frac{2m_f}{2m_f+1}(D-1+\frac{1}{2(m_0+m_\varepsilon)})}\right). \tag{82}$$

Similar to the proof in Appendix D.1, we can choose

$$N_0 = \frac{4n^2}{\eta_n(\tilde{\mathbf{K}})^2}, \tag{83}$$

and obtain that when $N \geq N_0$,

$$\|f_t - g_t\|_{L_2(P_{\mathbf{X}})} = O_{\mathbb{P}}\left(n^{-1/2}\right). \tag{84}$$

To bound the difference between the empirical norm $\|g_t - f^*\|_n$ and $\|g_t - f^*\|_{L_2(P_{\mathbf{X}})}$. By (95) and Lemma 30, we have

$$\|g_t\|^2_{\mathcal{N}_\sigma(\Omega)} \leq \sigma_n^{-2m_0}\|g_t\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)} \leq C_{17}\sigma_n^{-2m_0}\|\tilde{g}\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)} = O_{\mathbb{P}}\left(n^{\nu_1}(\log n)^{\nu_2}\right), \tag{85}$$

for some positive constants C_{17}, where

$$\begin{aligned}
\sigma_n \asymp& 1,\\
\nu_1 =& \frac{2(m_0+m_\varepsilon-m_f)}{2m_f+1},\\
\nu_2 =& 2(m_f-m_0-m_\varepsilon) + \frac{1}{2m_f+1}\left(\frac{m_f}{2(m_0+m_\varepsilon)}-1\right).
\end{aligned}$$

Consider function class $\mathcal{G} = \{h : h = (g_t - f^*)/(Cn^{\nu_1/2}(\log n)^{\nu_2/2}\}$, where the constant C is taken such that $\|h_1\|_{\mathcal{N}_\sigma(\Omega)} < 1$ for all $h_1 \in \mathcal{G}$. Select $r = \frac{1}{2}\frac{2m_f+1}{m_0+m_\varepsilon} > \frac{1}{2}\frac{1}{m_0+m_\varepsilon}$, then Lemma 31 leads to

$$\|h_1\|_{L_\infty(\Omega)} \leq C_1\|h_1\|^{1-\frac{2m_f+1}{2(m_0+m_\varepsilon)}}_{L_2(P_{\mathbf{X}})}\|h_1\|^{\frac{2m_f+1}{2(m_0+m_\varepsilon)}}_{\mathcal{N}_\sigma(\Omega)},$$

for some positive constants C_1 and all $h_1 \in \mathcal{G}$, which implies

$$c_1 := \sup_{h_1\in\mathcal{G}}\|h_1\|_{L_\infty(\Omega)} \leq C_2R_1^{1-\frac{2m_f+1}{2(m_0+m_\varepsilon)}},$$

where $R_1 = \sup_{h_1\in\mathcal{G}}\|h_1\|_{L_2(P_{\mathbf{X}})} \leq \sup_{h_1\in\mathcal{G}}\|h_1\|_{L_\infty(\Omega)} \leq \sup_{h_1\in\mathcal{G}}\|h_1\|_{\mathcal{N}_\sigma(\Omega)} < 1$, because of the reproducing property. Taking $c = C_2R_1^{1-\frac{1}{2(m_0+m_\varepsilon)}} < 1$, and we also let $\delta_n =$

$C_3(n^{-1}c^2)^{\frac{m_0+m_\varepsilon}{2(m_0+m_\varepsilon)+1}}(\log n)^{\frac{D-1}{2}+\frac{1}{4(m_0+m_\varepsilon)}}$, for some positive constants C_2 and C_3 in Lemma 23, it can be checked that

$$C_4 n\delta_n^2 c^{-2} \geq H(\delta, \mathcal{B}_{\mathcal{H}_\sigma(\Omega)}, \|\cdot\|_{L_\infty(\Omega)}),$$

for some positive constants C_4, which implies the conditions of Lemma 23 are fulfilled. Applying Lemma 23 to the case $\|g_t - f^*\|_{L_2(P_{\mathbf{X}})}^2 \geq \delta_n^2 n^{\nu_1}(\log n)^{\nu_2}$, together with (82), calculations similar to the proof in section D.1 shows

$$\|g_t - f^*\|_{L_2(P_{\mathbf{X}})} = O_{\mathbb{P}}\left(n^{-\frac{m_f}{2m_f+1}}(\log n)^{\frac{m_f}{2m_f+1}(D-1+\frac{1}{2(m_0+m_\varepsilon)})}\right).$$

This finishes the proof. ■

F.2 With Weight Decay

The results can be obtained by merely repeating the proof in Appendix D.2, where the only difference is that the corresponding convergence rate for I_2 (in Equation 62 of Appendix D.2) is obtained via the proof in Appendix F.1. Thus we omit it here.

Appendix G. Proof of Theorem 19

Similar to (29), we have

$$\tilde{g}_t(\mathbf{X}) = \left(\mathbf{I} - (\mathbf{I} - \beta\mathbf{K}_1)^t\right)\boldsymbol{y},$$

thus

$$\tilde{g}_t(\mathbf{X}) - f^*(\mathbf{X}) = -(\mathbf{I} - \beta\mathbf{K}_1)^t f^*(\mathbf{X}) + \left(\mathbf{I} - (\mathbf{I} - \beta\mathbf{K}_1)^t\right)\boldsymbol{\epsilon}, \tag{86}$$

where $\mathbf{K}_1 = (K_1(\boldsymbol{x}_j - \boldsymbol{x}_k))_{jk}$, and $f^*(\mathbf{X}) = (f^*(\boldsymbol{x}_1), ..., f^*(\boldsymbol{x}_n))^T$. Taking expectation with respect to $\boldsymbol{\epsilon}$, the mean squared prediction error of $\tilde{g}_t$ with respect to the empirical norm is given by

$$\begin{aligned}\mathbb{E}\|\tilde{g}_t - f^*\|_n^2 =& \frac{1}{n}\left((f^*(\mathbf{X}))^T(\mathbf{I} - \beta\mathbf{K}_1)^{2t}f^*(\mathbf{X}) + \sigma_\epsilon^2 \mathrm{tr}\left(\mathbf{I} - (\mathbf{I} - \beta\mathbf{K}_1)^t\right)^2\right)\\ =& \frac{1}{n}J_{11} + \frac{1}{n}J_{12}.\end{aligned} \tag{87}$$

By the representer theorem, the solution to (33) is given by

$$\tilde{g}(\boldsymbol{x}) = \mathbf{k}_1(\boldsymbol{x})^T(\mathbf{K}_1 + n\lambda\mathbf{I})^{-1}\boldsymbol{y}, \tag{88}$$

where $\mathbf{k}_1(\cdot) = (K_1(\cdot - \boldsymbol{x}_1), \ldots, K_1(\cdot - \boldsymbol{x}_n))^T$. Thus, the mean squared prediction error with respect to the empirical norm of $\tilde{g}$ can be computed by

$$\begin{aligned}&\mathbb{E}\|\tilde{g} - f^*\|_n^2\\ =& \frac{1}{n}\left((n\lambda)^2(f^*(\mathbf{X}))^T(\mathbf{K}_1 + n\lambda\mathbf{I})^{-2}f^*(\mathbf{X}) + \sigma_\epsilon^2 \mathrm{tr}\left((\mathbf{K}_1 + n\lambda\mathbf{I})^{-1}\mathbf{K}_1^2(\mathbf{K}_1 + n\lambda\mathbf{I})^{-1}\right)^2\right)\\ =& J_{21} + J_{22}.\end{aligned} \tag{89}$$

Let $\eta_1 \geq \ldots \geq \eta_n > 0$ and $\boldsymbol{v}_j$, $j = 1, \ldots, n$ be the eigenvalues and corresponding eigenvectors of $\mathbf{K}_1$, respectively. By the basic inequalities $1 - u \leq \exp(-u) \leq 2e(1+u)^{-2}$ for any $u > 0$, the term J_{11} can be bounded by

$$\begin{aligned}
J_{11} =& \sum_{j=1}^{n} (1 - \beta\eta_j)^{2t} (\boldsymbol{v}_j^T f^*(\mathbf{X}))^2 \leq \sum_{j=1}^{n} (1 - \beta\eta_j)^{t} (\boldsymbol{v}_j^T f^*(\mathbf{X}))^2 \\
\leq& \sum_{j=1}^{n} \exp(-\beta t\eta_j)(\boldsymbol{v}_j^T f^*(\mathbf{X}))^2 \leq 2e \sum_{j=1}^{n} \frac{(\beta t)^{-2}}{((\beta t)^{-1} + \eta_j)^2} (\boldsymbol{v}_j^T f^*(\mathbf{X}))^2 \\
=& 2e(\beta t)^{-2} (f^*(\mathbf{X}))^T (\mathbf{K}_1 + (\beta t)^{-1}\mathbf{I})^{-2} f^*(\mathbf{X}) \\
=& 2eJ_{21},
\end{aligned} \tag{90}$$

where the last equality is because we choose $n\lambda = (\beta t)^{-1}$.

Next, we consider J_{12}. Let r be the smallest integer such that $\beta t\eta_r \leq 1$. Then for $j = 1, ..., r-1$, we have

$$1 - (1 - \beta\eta_j)^t \leq 1 \leq \frac{2\beta t\eta_j}{1 + \beta t\eta_j}, \tag{91}$$

and for $j = r, ..., n$, we have

$$1 - (1 - \beta\eta_j)^t \leq \beta t\eta_j \leq \frac{2\beta t\eta_j}{1 + \beta t\eta_j}, \tag{92}$$

where the first inequality is by Bernoulli's inequality. Combining (91) and (92), we have

$$1 - (1 - \beta\eta_j)^t \leq \frac{2\beta t\eta_j}{1 + \beta t\eta_j}, \tag{93}$$

for all $j = 1, \ldots, n$. By (93), the second term J_{12} in (87) can be bounded by

$$\begin{aligned}
J_{12} =& \sigma_\epsilon^2 \sum_{j=1}^{n} (1 - (1 - \beta\eta_j)^t)^2 \leq \sigma_\epsilon^2 \sum_{j=1}^{n} \frac{4(\beta t\eta_j)^2}{(1 + \beta t\eta_j)^2} \\
=& 4\sigma_\epsilon^2 \mathrm{tr}\left((\mathbf{K}_1 + n\lambda\mathbf{I})^{-1}\mathbf{K}_1^2(\mathbf{K}_1 + n\lambda\mathbf{I})^{-1}\right)^2 = 4J_{22},
\end{aligned} \tag{94}$$

where in the second equality, we use $n\lambda = (\beta t)^{-1}$ again. By (87), (89) (90) and (94), and $2e > 4$, we have

$$\mathbb{E}\|g_t - f^*\|_n^2 \leq 2e\mathbb{E}\|\tilde{g} - f^*\|_n^2,$$

which finishes the proof of (34).

Next, we consider the RKHS norm of $\tilde{g}_t$ and show that (35) holds. Direct computation shows that

$$\begin{aligned}
&\|g_t\|^2_{\mathcal{H}_{K_1}(\Omega)} = g_t(\mathbf{X})^T \mathbf{K}_1^{-1} g_t(\mathbf{X}) = \sum_{j=1}^{n} \frac{(1 - (1 - \beta\eta_j)^t)^2}{\eta_j} (\boldsymbol{v}_j^T \boldsymbol{y})^2 \\
\leq& \sum_{j=1}^{n} \frac{4(\beta t)^2\eta_j}{(1 + \beta t\eta_j)^2} (\boldsymbol{v}_j^T \boldsymbol{y})^2 = 4\boldsymbol{y}^T (\mathbf{K}_1 + (\beta t)^{-1}\mathbf{I})^{-1} \mathbf{K}_1 (\mathbf{K}_1 + (\beta t)^{-1}\mathbf{I})^{-1} \boldsymbol{y} \\
=& 4\|\tilde{g}\|^2_{\mathcal{H}_{K_1}(\Omega)},
\end{aligned} \tag{95}$$

where the inequality is by (93), and the last equality is because $n\lambda = (\beta t)^{-1}$. This finishes the proof of (35). ■

Appendix H. Proof of Lemmas in Appendix B

H.1 Proof of Lemma 16

From Assumption 2 or Assumption 3, for any $\boldsymbol{x}, \boldsymbol{x}' \in \Omega$, the Fourier inversion theorem yields

$$
\begin{aligned}
&\left|\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}'}\big(K(\boldsymbol{x}+\boldsymbol{\varepsilon}-(\boldsymbol{x}'+\boldsymbol{\varepsilon}'))\big) - \frac{1}{N^2}\sum_{k=1}^N\sum_{j=1}^N K(\boldsymbol{x}+\boldsymbol{\varepsilon}_j-(\boldsymbol{x}'+\boldsymbol{\varepsilon}_k))\right| \\
=&\left|\int_{\mathbb{R}^D}\mathbb{E}_{\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}'}\big(e^{i\boldsymbol{\omega}^T(\boldsymbol{x}+\boldsymbol{\varepsilon}-\boldsymbol{x}'-\boldsymbol{\varepsilon}')}\big)\mathcal{F}(K)(\boldsymbol{\omega}) - \frac{1}{N^2}\sum_{k=1}^N\sum_{j=1}^N e^{i\boldsymbol{\omega}^T(\boldsymbol{x}+\boldsymbol{\varepsilon}_k-\boldsymbol{x}'-\boldsymbol{\varepsilon}_j')}\mathcal{F}(K)(\boldsymbol{\omega})\mathrm{d}\boldsymbol{\omega}\right| \\
\leq&\int_{\mathbb{R}^D}\left|\big|\mathbb{E}_{\boldsymbol{\varepsilon}}\big(e^{i\boldsymbol{\omega}^T\boldsymbol{\varepsilon}}\big)\big|^2 - \left|\frac{1}{N}\sum_{k=1}^N e^{i\boldsymbol{\omega}^T\boldsymbol{\varepsilon}_k}\right|^2\right|\mathcal{F}(K)(\boldsymbol{\omega})\mathrm{d}\boldsymbol{\omega} \\
\leq&\int_{\mathbb{R}^D}\left|\mathbb{E}_{\boldsymbol{\varepsilon}}\big(e^{i\boldsymbol{\omega}^T\boldsymbol{\varepsilon}}\big) - \frac{1}{N}\sum_{k=1}^N e^{i\boldsymbol{\omega}^T\boldsymbol{\varepsilon}_k}\right|\left(\big|\mathbb{E}_{\boldsymbol{\varepsilon}}\big(e^{-i\boldsymbol{\omega}^T\boldsymbol{\varepsilon}}\big)\big| + \left|\frac{1}{N}\sum_{k=1}^N e^{-i\boldsymbol{\omega}^T\boldsymbol{\varepsilon}_k}\right|\right)\mathcal{F}(K)(\boldsymbol{\omega})\mathrm{d}\boldsymbol{\omega} \\
\leq&2\int_{\mathbb{R}^D}\left|\mathbb{E}_{\boldsymbol{\varepsilon}}\big(e^{i\boldsymbol{\omega}^T\boldsymbol{\varepsilon}}\big) - \frac{1}{N}\sum_{k=1}^N e^{i\boldsymbol{\omega}^T\boldsymbol{\varepsilon}_k}\right|\mathcal{F}(K)(\boldsymbol{\omega})\mathrm{d}\boldsymbol{\omega}. \qquad (96)
\end{aligned}
$$

According to Assumption 4, $\boldsymbol{\varepsilon}$ is sub-Gaussian. From Csörgő (1985), we can have the following error estimate for the empirical characteristic function $\frac{1}{N}\sum_{k=1}^N e^{-i\boldsymbol{\omega}^T\boldsymbol{\varepsilon}_k}$ almost surely. Specifically, for any $A > 0$, we have

$$
\limsup_{N\to\infty}\sqrt{\frac{N}{\log N}}\sup_{\|\omega\|_2\leq N^A}\left|\mathbb{E}_{\boldsymbol{\varepsilon}}\big(e^{i\boldsymbol{\omega}^T\boldsymbol{\varepsilon}}\big) - \frac{1}{N}\sum_{k=1}^N e^{i\boldsymbol{\omega}^T\boldsymbol{\varepsilon}_k}\right| \leq 2+\sqrt{2\min(A,1)}+4\sqrt{1+(A+\frac{1}{2})D}. \qquad (97)
$$

By (97), (96) can be further bounded by

$$
\begin{aligned}
&2\int_{\mathbb{R}^D}\left|\mathbb{E}_{\boldsymbol{\varepsilon}}\big(e^{i\boldsymbol{\omega}^T\boldsymbol{\varepsilon}}\big) - \frac{1}{N}\sum_{k=1}^N e^{i\boldsymbol{\omega}^T\boldsymbol{\varepsilon}_k}\right|\mathcal{F}(K)(\boldsymbol{\omega})\mathrm{d}\boldsymbol{\omega} \\
=&2\int_{\|\boldsymbol{\omega}\|_2\leq N^A}\left|\mathbb{E}_{\boldsymbol{\varepsilon}}\big(e^{i\boldsymbol{\omega}^T\boldsymbol{\varepsilon}}\big) - \frac{1}{N}\sum_{k=1}^N e^{i\boldsymbol{\omega}^T\boldsymbol{\varepsilon}_k}\right|\mathcal{F}(K)(\boldsymbol{\omega})\mathrm{d}\boldsymbol{\omega} \\
&+\int_{\|\boldsymbol{\omega}\|_2> N^A}\left|\mathbb{E}_{\boldsymbol{\varepsilon}}\big(e^{i\boldsymbol{\omega}^T\boldsymbol{\varepsilon}}\big) - \frac{1}{N}\sum_{k=1}^N e^{i\boldsymbol{\omega}^T\boldsymbol{\varepsilon}_k}\right|\mathcal{F}(K)(\boldsymbol{\omega})\mathrm{d}\boldsymbol{\omega} \\
=&O_{\mathbb{P}}\bigg(\int_{\|\boldsymbol{\omega}\|_2\leq N^A}\sqrt{\frac{\log N}{N}}\mathcal{F}(K)(\boldsymbol{\omega})\mathrm{d}\boldsymbol{\omega} + 2\int_{\|\boldsymbol{\omega}\|_2> N^A}\mathcal{F}(K)(\boldsymbol{\omega})\mathrm{d}\boldsymbol{\omega}\bigg).
\end{aligned}
$$

If Assumption 2 is satisfied, then we can set $A = (2m_0 - d)^{-1}$ and obtain

$$
\begin{aligned}
&\int_{\|\boldsymbol{\omega}\|_2 \le N^A} \sqrt{\frac{\log N}{N}} \mathcal{F}(K)(\boldsymbol{\omega}) \mathrm{d}\boldsymbol{\omega} + 2\int_{\|\boldsymbol{\omega}\|_2 > N^A} \mathcal{F}(K)(\boldsymbol{\omega}) \mathrm{d}\boldsymbol{\omega} \\
\le & C_1 \Big(\int_{\|\boldsymbol{\omega}\|_2 \le N^A} \sqrt{\frac{\log N}{N}} (1 + \|\boldsymbol{\omega}\|_2^2)^{-m_0} \mathrm{d}\boldsymbol{\omega} + 2\int_{\|\boldsymbol{\omega}\|_2 > N^A} (1 + \|\boldsymbol{\omega}\|_2^2)^{-m_0} \mathrm{d}\boldsymbol{\omega} \Big) \\
\lesssim & \sqrt{\frac{\log N}{N}},
\end{aligned}
$$

for some positive constants C_1, where the last inequality is because $m_0 > D/2$. Similarly, if Assumption 3 is satisfied, then we set $A = (2m_0 - 1)^{-1}$ and get

$$
\begin{aligned}
&\int_{\|\boldsymbol{\omega}\|_2 \le N^A} \sqrt{\frac{\log N}{N}} \mathcal{F}(K)(\boldsymbol{\omega}) \mathrm{d}\boldsymbol{\omega} + 2\int_{\|\boldsymbol{\omega}\|_2 > N^A} \mathcal{F}(K)(\boldsymbol{\omega}) \mathrm{d}\boldsymbol{\omega} \\
\le & C_2 \Big(\int_{\|\boldsymbol{\omega}\|_2 \le N^A} \sqrt{\frac{\log N}{N}} \prod_{j=1}^{D} (1 + \omega_j^2)^{-m_0} \mathrm{d}\boldsymbol{\omega} + 2\int_{\|\boldsymbol{\omega}\|_2 > N^A} \prod_{j=1}^{D} (1 + \omega_j^2)^{-m_0} \mathrm{d}\boldsymbol{\omega} \Big) \\
\lesssim & \sqrt{\frac{\log N}{N}} + \int_{\max_j |\omega_j| \ge N^A/\sqrt{D}} \prod_{j=1}^{D} (1 + \omega_j^2)^{-m_0} \mathrm{d}\boldsymbol{\omega} \\
\lesssim & \sqrt{\frac{\log N}{N}},
\end{aligned}
$$

for some positive constants C_2.
This finishes the proof. ■

H.2 Proof of Lemma 17

For any $\boldsymbol{x} \in \Omega$, by (26) and (30), we have

$$
\begin{aligned}
f_{\mathrm{t}}(\boldsymbol{x}) &= \mathbf{k}(\boldsymbol{x})^T (\alpha/\beta \mathbf{I} + \mathbf{K})^{-1} \boldsymbol{y} - \beta \mathbf{k}(\boldsymbol{x})^T ((1-\alpha)\mathbf{I} - \beta \mathbf{K})^{\mathrm{t}} (\alpha \mathbf{I} + \beta \mathbf{K})^{-1} \boldsymbol{y}, \\
g_{\mathrm{t}}(\boldsymbol{x}) &= \tilde{\mathbf{k}}(\boldsymbol{x})^T (\alpha/\beta \mathbf{I} + \tilde{\mathbf{K}})^{-1} \boldsymbol{y} - \beta \tilde{\mathbf{k}}(\boldsymbol{x})^T ((1-\alpha)\mathbf{I} - \beta \tilde{\mathbf{K}})^{\mathrm{t}} (\alpha \mathbf{I} + \beta \tilde{\mathbf{K}})^{-1} \boldsymbol{y}.
\end{aligned}
$$

Applying the triangle inequality yields

$$
\begin{aligned}
& \|f_{\mathrm{t}} - g_{\mathrm{t}}\|_{L_\infty(\Omega)} \\
\le & \|\mathbf{k}(\cdot)^T (\alpha/\beta \mathbf{I} + \mathbf{K})^{-1} \boldsymbol{y} - \tilde{\mathbf{k}}(\cdot)^T (\alpha/\beta \mathbf{I} + \tilde{\mathbf{K}})^{-1} \boldsymbol{y}\|_{L_\infty(\Omega)} \\
& + \|\mathbf{k}(\cdot)^T ((1-\alpha)\mathbf{I} - \beta \mathbf{K})^{\mathrm{t}} (\alpha/\beta \mathbf{I} + \mathbf{K})^{-1} \boldsymbol{y} - \tilde{\mathbf{k}}(\cdot)^T ((1-\alpha)\mathbf{I} - \tilde{\mathbf{K}})^{\mathrm{t}} (\alpha/\beta \mathbf{I} + \tilde{\mathbf{K}})^{-1} \boldsymbol{y}\|_{L_\infty(\Omega)} \\
\le & \|\big(\mathbf{k}(\cdot) - \tilde{\mathbf{k}}(\cdot)\big)^T (\alpha/\beta \mathbf{I} + \mathbf{K})^{-1} \boldsymbol{y}\|_{L_\infty(\Omega)} && (98) \\
& + \|\tilde{\mathbf{k}}(\cdot)^T \big((\alpha/\beta \mathbf{I} + \mathbf{K})^{-1} - (\alpha/\beta \mathbf{I} + \tilde{\mathbf{K}})^{-1}\big) \boldsymbol{y}\|_{L_\infty(\Omega)} && (99) \\
& + \|\big(\mathbf{k}(\cdot) - \tilde{\mathbf{k}}(\cdot)\big)^T ((1-\alpha)\mathbf{I} - \beta \mathbf{K})^{\mathrm{t}} (\alpha/\beta \mathbf{I} + \mathbf{K})^{-1} \boldsymbol{y}\|_{L_\infty(\Omega)} && (100) \\
& + \|\tilde{\mathbf{k}}(\cdot)^T ((1-\alpha)\mathbf{I} - \beta \mathbf{K})^{\mathrm{t}} \big((\alpha/\beta \mathbf{I} + \tilde{\mathbf{K}})^{-1} - (\alpha/\beta \mathbf{I} + \mathbf{K})^{-1}\big) \boldsymbol{y}\|_{L_\infty(\Omega)} && (101) \\
& + \|\tilde{\mathbf{k}}(\cdot)^T \big(((1-\alpha)\mathbf{I} - \beta \tilde{\mathbf{K}})^{\mathrm{t}} - ((1-\alpha)\mathbf{I} - \beta \mathbf{K})^{\mathrm{t}}\big) (\alpha/\beta \mathbf{I} + \tilde{\mathbf{K}})^{-1} \boldsymbol{y}\|_{L_\infty(\Omega)}. && (102)
\end{aligned}
$$

For (98), we have

$$
\begin{aligned}
&\|\big(\mathbf{k}(\cdot)-\tilde{\mathbf{k}}(\cdot)\big)^T(\alpha/\beta\mathbf{I}+\mathbf{K})^{-1}\boldsymbol{y}\|_{L_\infty(\Omega)}\\
\leq&\eta_n(\alpha/\beta\mathbf{I}+\mathbf{K})^{-1}\|\boldsymbol{y}\|_2\left(\sup_{\boldsymbol{x}\in\Omega}\|\mathbf{k}(\boldsymbol{x})-\tilde{\mathbf{k}}(\boldsymbol{x})\|_2\right)\\
\leq&\eta_n(\mathbf{K})^{-1}\sqrt{\sum_{j=1}^n y_j^2}\left(\sup_{\boldsymbol{x}\in\Omega}\sqrt{\sum_{j=1}^n\big(K_s(\boldsymbol{x}_i,\boldsymbol{x})-\tilde{K}_s(\boldsymbol{x}_i,\boldsymbol{x})\big)^2}\right)\\
=&\eta_n(\mathbf{K})^{-1}\sqrt{\sum_{j=1}^n y_j^2}\,O_{\mathbb{P}}\Bigg(\sqrt{\frac{n\log N}{N}}\Bigg)\\
\leq&\eta_n(\mathbf{K})^{-1}\sqrt{3n}\big(\max_{j=1,\ldots,n}|f^*(\boldsymbol{x}_j)|+\sqrt{\frac{1}{n}\sum_{j=1}^n|\epsilon_j|^2}\big)O_{\mathbb{P}}\Bigg(\sqrt{\frac{n\log N}{N}}\Bigg)\\
=&O_{\mathbb{P}}\left(\eta_n(\mathbf{K})^{-1}n\sqrt{\frac{\log N}{N}}\right)\\
=&O_{\mathbb{P}}\left(\eta_n(\tilde{\mathbf{K}})^{-1}n\sqrt{\frac{\log N}{N}}\right),
\end{aligned}
\tag{103}
$$

where the fourth line is by Lemma 16, the sixth line is because $\max_{j=1,\ldots,n}|f^*(\boldsymbol{x}_j)|\lesssim\|f^*\|_{\mathcal{W}^{m_f}(\Omega_1)}$ and ϵ_j's are sub-Gaussian variables, and the last line is because

$$
\begin{aligned}
\eta_n(\mathbf{K})=&\eta_n(\tilde{\mathbf{K}}+(\mathbf{K}-\tilde{\mathbf{K}}))\geq\eta_n(\tilde{\mathbf{K}})-n\max_{j,k}|K_S(\boldsymbol{x}_j,\boldsymbol{x}_k)-\tilde{K}_S(\boldsymbol{x}_j,\boldsymbol{x}_k)|\\
\geq&\eta_n(\tilde{\mathbf{K}})-n\sqrt{\frac{\log N}{N}}\geq\frac{1}{2}\eta_n(\tilde{\mathbf{K}}).
\end{aligned}
\tag{104}
$$

By Gershgorin's theorem (Varga, 2010), we have

$$
\|\mathbf{K}-\tilde{\mathbf{K}}\|_2\leq n\max_{j,k}|K_S(\boldsymbol{x}_j,\boldsymbol{x}_k)-\tilde{K}_S(\boldsymbol{x}_j,\boldsymbol{x}_k)|=O_{\mathbb{P}}\left(n\sqrt{\frac{\log N}{N}}\right).
\tag{105}
$$

Therefore, it can be checked that

$$
\begin{aligned}
&\|(\alpha/\beta\mathbf{I}+\mathbf{K})^{-1}-(\alpha/\beta\mathbf{I}+\tilde{\mathbf{K}})^{-1}\|_2=\|(\alpha/\beta\mathbf{I}+\mathbf{K})^{-1}(\alpha/\beta\mathbf{I}+\tilde{\mathbf{K}})^{-1}(\mathbf{K}-\tilde{\mathbf{K}})\|_2\\
\leq&\frac{n\max_{j,k}|K_S(\boldsymbol{x}_j,\boldsymbol{x}_k)-\tilde{K}_S(\boldsymbol{x}_j,\boldsymbol{x}_k)|}{\eta_n(\mathbf{K})\eta_n(\tilde{\mathbf{K}})}=O_{\mathbb{P}}\bigg(\frac{n\sqrt{\log N/N}}{\eta_n(\mathbf{K})\eta_n(\tilde{\mathbf{K}})}\bigg)\\
=&O_{\mathbb{P}}\bigg(\frac{n\sqrt{\log N/N}}{\eta_n(\tilde{\mathbf{K}})^2}\bigg),
\end{aligned}
\tag{106}
$$

where second line is because of Gershgorin's theorem (Varga, 2010), the third line is from Lemma 16, and the last line is from (104). Therefore, plugging (106) into (99) gives us

$$\begin{aligned}
&\|\tilde{\mathbf{k}}(\cdot)^T\big((\alpha/\beta\mathbf{I}+\mathbf{K})^{-1}-(\alpha/\beta\mathbf{I}+\tilde{\mathbf{K}})^{-1}\big)\boldsymbol{y}\|_{L_\infty(\Omega)}\\
\le&\sup_{\boldsymbol{x}\in\Omega}\|\tilde{\mathbf{k}}(\boldsymbol{x})\|_2\|\mathbf{y}\|_2\|(\alpha/\beta\mathbf{I}+\mathbf{K})^{-1}-(\alpha/\beta\mathbf{I}+\tilde{\mathbf{K}})^{-1}\|_2\\
\le& n\left(\sup_{\boldsymbol{x}\in\Omega}\max_{j=1,\dots,n}\tilde{K}_s(\boldsymbol{x}_j,\boldsymbol{x})\right)\left(\sqrt{3}\max_{j=1,\dots,n}|f^*(\boldsymbol{x}_j)|+\sqrt{3}\sqrt{\frac{1}{n}\sum_{j=1}^n|\varepsilon_j|^2}\right)O_{\mathbb{P}}\left(\frac{n\sqrt{\log N/N}}{\eta_n(\tilde{\mathbf{K}})^2}\right)\\
=&O_{\mathbb{P}}\left(\frac{n^2\sqrt{\log N/N}}{\eta_n(\tilde{\mathbf{K}})^2}\right).
\end{aligned}\tag{107}$$

For (100), because $0<1-\alpha-\beta\eta_1(\mathbf{K})<1$, we have

$$\begin{aligned}
&\|\big(\mathbf{k}(\cdot)-\tilde{\mathbf{k}}(\cdot)\big)^T((1-\alpha)\mathbf{I}-\beta\mathbf{K})^{\mathrm{t}}(\alpha/\beta\mathbf{I}+\mathbf{K})^{-1}\boldsymbol{y}\|_{L_\infty(\Omega)}\\
\le&\eta_n(\mathbf{K})^{-1}\|\boldsymbol{y}\|_2\left(\sup_{\boldsymbol{x}\in\Omega}\|\mathbf{k}(\boldsymbol{x})-\tilde{\mathbf{k}}(\boldsymbol{x})\|_2\right)\\
=&O_{\mathbb{P}}\left(\eta_n(\mathbf{K})^{-1}n\sqrt{\frac{\log N}{N}}\right)=O_{\mathbb{P}}\left(\eta_n(\tilde{\mathbf{K}})^{-1}n\sqrt{\frac{\log N}{N}}\right),
\end{aligned}\tag{108}$$

where the last line is from (103) and (104).

Similarly, for (101), we have

$$\begin{aligned}
&\|\tilde{\mathbf{k}}(\cdot)^T((1-\alpha)\mathbf{I}-\beta\mathbf{K})^{\mathrm{t}}\big((\alpha/\beta\mathbf{I}+\tilde{\mathbf{K}})^{-1}-(\alpha/\beta\mathbf{I}+\mathbf{K})^{-1}\big)\boldsymbol{y}\|_{L_\infty(\Omega)}\\
\le&\sup_{\boldsymbol{x}\in\Omega}\|\tilde{\mathbf{k}}(\boldsymbol{x})\|_2\|\mathbf{y}\|_2\|(\alpha/\beta\mathbf{I}+\mathbf{K})^{-1}-(\alpha/\beta\mathbf{I}+\tilde{\mathbf{K}})^{-1}\|_2\\
=&O_{\mathbb{P}}\left(\frac{n^2\sqrt{\log N/N}}{\eta_n(\tilde{\mathbf{K}})^2}\right).
\end{aligned}\tag{109}$$

For (102), we have

$$\begin{aligned}
&\|\tilde{\mathbf{k}}(\cdot)^T\big(((1-\alpha)\mathbf{I}-\beta\tilde{\mathbf{K}})^{\mathrm{t}}-((1-\alpha)\mathbf{I}-\beta\mathbf{K})^{\mathrm{t}}\big)(\alpha/\beta\mathbf{I}+\tilde{\mathbf{K}})^{-1}\boldsymbol{y}\|_{L_\infty(\Omega)}\\
\le&\sup_{\boldsymbol{x}\in\Omega}\|\tilde{\mathbf{k}}(\boldsymbol{x})\|_2\|(\alpha/\beta\mathbf{I}+\tilde{\mathbf{K}})^{-1}\|_2\|\boldsymbol{y}\|_2\|((1-\alpha)\mathbf{I}-\beta\tilde{\mathbf{K}})^{\mathrm{t}}-((1-\alpha)\mathbf{I}-\beta\mathbf{K})^{\mathrm{t}}\|_2\\
\le&\frac{n}{\eta_n(\tilde{\mathbf{K}})}\|((1-\alpha)\mathbf{I}-\beta\tilde{\mathbf{K}})^{\mathrm{t}}-((1-\alpha)\mathbf{I}-\beta\mathbf{K})^{\mathrm{t}}\|_2.
\end{aligned}\tag{110}$$

The term $\|((1-\alpha)\mathbf{I}-\beta\tilde{\mathbf{K}})^{\mathsf{t}}-((1-\alpha)\mathbf{I}-\beta\mathbf{K})^{\mathsf{t}}\|_2$ can be further bounded by

$$
\begin{aligned}
&\|((1-\alpha)\mathbf{I}-\beta\tilde{\mathbf{K}})^{\mathsf{t}}-((1-\alpha)\mathbf{I}-\beta\mathbf{K})^{\mathsf{t}}\|_2\\
\leq&\|\beta\tilde{\mathbf{K}}-\beta\mathbf{K}\|_2\left\|\sum_{j=0}^{\mathsf{t}-1}\left((1-\alpha)\mathbf{I}-\beta\tilde{\mathbf{K}}\right)^j\left((1-\alpha)\mathbf{I}-\beta\mathbf{K}\right)^{\mathsf{t}-1-j}\right\|_2\\
=&O_{\mathbb{P}}\left(\beta n\sqrt{\frac{\log N}{N}}\right)\left(\sum_{j=0}^{\mathsf{t}-1}\left|(1-\alpha)-\beta\eta_n(\tilde{\mathbf{K}})\right|^j\left|(1-\alpha)-\beta\eta_n(\mathbf{K})\right|^{\mathsf{t}-1-j}\right)\\
\leq&O_{\mathbb{P}}\left(\sqrt{\frac{\log N}{N}}\right)\left(\sum_{j=0}^{\mathsf{t}-1}\left|(1-\alpha)-\beta(\eta_n(\mathbf{K})-\eta_1(\tilde{\mathbf{K}}-\mathbf{K}))\right|^j\left|(1-\alpha)-\beta\eta_n(\mathbf{K})\right|^{\mathsf{t}-1-j}\right)\\
\leq&O_{\mathbb{P}}\left(\sqrt{\frac{\log N}{N}}\right)\left(\sum_{j=0}^{\mathsf{t}-1}\left|(1-\alpha)-\beta\eta_n(\mathbf{K})+O_{\mathbb{P}}\left(n\sqrt{\frac{\log N}{N}}\right)\right|^j\left|(1-\alpha)-\beta\eta_n(\mathbf{K})\right|^{\mathsf{t}-1-j}\right)\\
\leq&O_{\mathbb{P}}\left(\mathsf{t}\sqrt{\frac{\log N}{N}}\left|1-\alpha-\beta\eta_n(\mathbf{K})+n\sqrt{\frac{\log N}{N}}\right|^{\mathsf{t}}\right),
\end{aligned}
\tag{111}
$$

where the second line is because of the basic identity $a^{\mathsf{t}}-b^{\mathsf{t}}=(a-b)(\sum_{j=0}^{\mathsf{t}-1}a^jb^{\mathsf{t}-1-j})$, the third line is because of (105), and the fifth line is by the second inequality in (104).

Since α,β, and $\eta_n(\mathbf{K})$ are not depending on N, we can let N_0 satisfy $n\sqrt{\frac{\log N_0}{N_0}}\leq(\alpha+\beta\eta_n(\mathbf{K}))/2$ such that for all $N>N_0+3$

$$
\left|1-\alpha-\beta\eta_n(\mathbf{K})+n\sqrt{\frac{\log N}{N}}\right|\leq\left|1-\frac{\alpha+\beta\eta_n(\mathbf{K})}{2}\right|.
$$

Let $\mathsf{t}_0=2/(\alpha+\beta\eta_n(\mathbf{K}))$, and $h(\mathsf{t})=\mathsf{t}(1-(\alpha+\beta\eta_n(\mathbf{K}))/2)^{\mathsf{t}}$. Basic calculation shows that if $\mathsf{t}>\mathsf{t}_0$, $h(\mathsf{t})$ is a decreasing function. Thus, $h(\mathsf{t})\leq h(\mathsf{t}_0)$. By the basic inequality $(1-x)^x\leq e^{-1}$, we obtain that if $\mathsf{t}>\mathsf{t}_0$, (111) can be further bounded by

$$
\begin{aligned}
&\mathsf{t}\sqrt{\frac{\log N}{N}}\left|1-\alpha-\beta\eta_n(\mathbf{K})+n\sqrt{\frac{\log N}{N}}\right|^{\mathsf{t}}\\
\leq&\sqrt{\frac{\log N}{N}}\mathsf{t}_0e^{-1}\leq\sqrt{\frac{\log N}{N}}\mathsf{t}_0\\
=&\sqrt{\frac{\log N}{N}}\frac{2}{\alpha+\beta\eta_n(\mathbf{K})}\leq\sqrt{\frac{\log N}{N}}\frac{2n}{n\beta\eta_n(\mathbf{K})}\\
\leq&C_1\frac{n\sqrt{\log N/N}}{\eta_n(\tilde{\mathbf{K}})},
\end{aligned}
\tag{112}
$$

for some positive constants C_1, where we use $n\beta$ is a constant. If $\mathfrak{t} \leq \mathfrak{t}_0$, then

$$\begin{aligned}
&\mathfrak{t}\sqrt{\frac{\log N}{N}}\left|1-\alpha-\beta\eta_n(\mathbf{K})+n\sqrt{\frac{\log N}{N}}\right|^{\mathfrak{t}}\\
\leq&\sqrt{\frac{\log N}{N}}\mathfrak{t}_0 \leq C_1\frac{n\sqrt{\log N/N}}{\eta_n(\tilde{\mathbf{K}})}, \qquad (113)
\end{aligned}$$

since $1-\alpha-\beta\eta_n(\mathbf{K})+n\sqrt{\frac{\log N}{N}} < 1$. Therefore, as long as $N > N_0+3$, by plugging (111), (112), and (113) in (110), we have

$$\begin{aligned}
&\|\tilde{\mathbf{k}}(\cdot)^T(\alpha/\beta\mathbf{I}+\tilde{\mathbf{K}})^{-1}\boldsymbol{y}\|_{L_\infty(\Omega)}\|\sum_{j=0}^{\mathfrak{t}-1}\big((1-\alpha)\mathbf{I}-\beta\tilde{\mathbf{K}}\big)^j\big((1-\alpha)\mathbf{I}-\beta\mathbf{K}\big)^{\mathfrak{t}-1-j}\|_2\\
=&O_{\mathbb{P}}\left(\frac{n^2\sqrt{\log N/N}}{\eta_n(\tilde{\mathbf{K}})^2}\right). \qquad (114)
\end{aligned}$$

Putting together (103), (107), (108), (109), and (114), we obtain the final result. ∎

H.3 Proof of Lemma 18

If Assumption 2 is satisfied, the Fourier inversion theorem implies that for any $\boldsymbol{x} \in \mathbb{R}^D$, it holds that

$$\begin{aligned}
\tilde{K}_S(\boldsymbol{x}) =& \int_{\mathbb{R}^D}\int_{\mathbb{R}^D} K(\boldsymbol{x}+\boldsymbol{\varepsilon}-\boldsymbol{\varepsilon}')p_\varepsilon(\boldsymbol{\varepsilon})p_\varepsilon(\boldsymbol{\varepsilon}')\mathrm{d}\boldsymbol{\varepsilon}\mathrm{d}\boldsymbol{\varepsilon}'\\
=&(2\pi)^{-D/2}\int_{\mathbb{R}^D}\int_{\mathbb{R}^D}\int_{\mathbb{R}^D} e^{-i(\boldsymbol{x}+\boldsymbol{\varepsilon}-\boldsymbol{\varepsilon}')^T\boldsymbol{\omega}}\mathcal{F}(K)(\boldsymbol{\omega})\mathrm{d}\boldsymbol{\omega}p_\varepsilon(\boldsymbol{\varepsilon})p_\varepsilon(\boldsymbol{\varepsilon}')\mathrm{d}\boldsymbol{\varepsilon}\mathrm{d}\boldsymbol{\varepsilon}'\\
=&(2\pi)^{-D/2}\int_{\mathbb{R}^D} e^{-i\boldsymbol{x}^T\boldsymbol{\omega}}\mathcal{F}(K)(\boldsymbol{\omega})|\varphi_\varepsilon(\boldsymbol{\omega})|^2\mathrm{d}\boldsymbol{\omega},
\end{aligned}$$

where φ_ε is the characteristic function of p_ε. Thus, by the Fourier theorem,

$$\mathcal{F}(\tilde{K}_S)(\boldsymbol{\omega}) = \mathcal{F}(K)(\boldsymbol{\omega})|\varphi_\varepsilon(\boldsymbol{\omega})|^2.$$

Therefore, for any $\boldsymbol{a} \in \mathbb{R}^n$, we have

$$\begin{aligned}
\boldsymbol{a}^T\tilde{\mathbf{K}}\boldsymbol{a} =& \sum_{j=1}^n\sum_{k=1}^n a_j\tilde{K}_S(\boldsymbol{x}_j-\boldsymbol{x}_k)a_k\\
=&(2\pi)^{-D/2}\int_{\mathbb{R}^D}\sum_{j,k=1}^n a_j e^{-i(\boldsymbol{x}_j-\boldsymbol{x}_k)^T\boldsymbol{\omega}}a_k\mathcal{F}(K)(\boldsymbol{\omega})|\varphi_\varepsilon(\boldsymbol{\omega})|^2\mathrm{d}\boldsymbol{\omega}\\
\geq&C_1\int_{\mathbb{R}^D}\left|\sum_{k=1} a_k e^{i\boldsymbol{\omega}^T\boldsymbol{x}_k}\right|^2\big(1+\|\boldsymbol{\omega}\|_2^2\big)^{-m_0}|\varphi_\varepsilon(\boldsymbol{\omega})|^2\mathrm{d}\boldsymbol{\omega}. \qquad (115)
\end{aligned}$$

where C_1 is a constant only depending on D. Similarly, if Assumption 3 and Assumption 4 (C2) are satisfied, the Fourier inversion theorem implies that for any $\boldsymbol{x} \in \mathbb{R}^D$,

$$\tilde{K}_S(\boldsymbol{x}, \boldsymbol{x}') =(2\pi)^{-D/2} \int_{\mathbb{R}^D} e^{-i(\boldsymbol{x}-\boldsymbol{x}')^T \boldsymbol{\omega}} \prod_{j=1}^{D} \mathcal{F}(K_j)(\omega_j) |\varphi_\varepsilon(\boldsymbol{\omega})|^2 \mathrm{d}\boldsymbol{\omega}.$$

Thus, for any $\{a_i\}_{i=1}^n \subset \mathbb{R}$, we have

$$\begin{aligned} \boldsymbol{a}^T \tilde{\mathbf{K}} \boldsymbol{a} &= \sum_{k,j=1}^{n} a_k \tilde{K}_S(\boldsymbol{x}_k, \boldsymbol{x}_j) a_j \\ &\geq C_2 \int_{\mathbb{R}^D} \left| \sum_{k=1} a_k e^{i\boldsymbol{\omega}^T \boldsymbol{x}_k} \right|^2 \prod_{j=1}^{D} |1+\omega_j^2|^{-m_0} |1+\sigma_n^2 \omega_j^2|^{-m_\varepsilon} \mathrm{d}\boldsymbol{\omega} \\ &\geq C_2 \int_{\mathbb{R}^D} \left| \sum_{k=1} a_k e^{i\boldsymbol{\omega}^T \boldsymbol{x}_k} \right|^2 (1+\|\boldsymbol{\omega}\|_2^2)^{-m_0 D} (1+\sigma_n^2 \|\boldsymbol{\omega}\|_2^2)^{-m_\varepsilon D} \mathrm{d}\boldsymbol{\omega}, \end{aligned} \tag{116}$$

where the positive constant C_2 is only depending on D.

We then apply Theorem 12.3 of Wendland (2004) on (115) and (116), respectively, and the final results can be straightforwardly derived. ■

Appendix I. Proof of Lemmas in Appendix D

In this section, we present the proof of lemmas in Appendix D.

I.1 Proof of Lemma 21

Let $\tilde{f}_n^*$ be the solution to the optimization problem

$$\min_{g \in \mathcal{H}_{\tilde{K}_S}(\mathbb{R}^D)} \|f^* - g\|_{L_2(\mathbb{R}^D)}^2 + \lambda_n \|g\|_{\mathcal{H}_{\tilde{K}_S}(\mathbb{R}^D)}^2. \tag{117}$$

Since f_n^* is the solution to (36), we have

$$\|f^* - f_n^*\|_{L_2(P_{\mathbf{X}})}^2 + \lambda_n \|f_n^*\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)}^2 \leq \|f^* - \tilde{f}_n^*\|_{L_2(P_{\mathbf{X}})}^2 + \lambda_n \|\tilde{f}_n^*\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)}^2. \tag{118}$$

Let $f_1 = f^* - \tilde{f}_n^*$. Then f_1 is well-defined in $\mathbb{R}^D$ and the Fourier inversion theorem implies that

$$\begin{aligned}
\|f_1\|_{L_2(P_{\mathbf{X}})}^2 =& \left(\int_\Omega \left|\int_{\mathbb{R}^D} e^{i\boldsymbol{x}^T\boldsymbol{\omega}}(\mathcal{F}(f_1)(\boldsymbol{\omega}))\mathrm{d}\boldsymbol{\omega}\right|^2 \mathrm{d}P_{\mathbf{X}}\right)\\
\leq& \left(\int_{\mathbb{R}^D}\left(\int_\Omega \left|e^{i\boldsymbol{x}^T\boldsymbol{\omega}}(\mathcal{F}(f_1)(\boldsymbol{\omega}))\right|^2 \mathrm{d}P_{\mathbf{X}}\right)^{1/2}\mathrm{d}\boldsymbol{\omega}\right)^2\\
\leq& C_1\left(\int_{\mathbb{R}^D}|(\mathcal{F}(f_1)(\boldsymbol{\omega}))|\,\mathrm{d}\boldsymbol{\omega}\right)^2\\
\leq& C_1\int_{\mathbb{R}^D}|(\mathcal{F}(f_1)(\boldsymbol{\omega}))|^2\,\mathrm{d}\boldsymbol{\omega}\\
=& C_1\|f_1\|_{L_2(\mathbb{R}^D)},
\end{aligned} \tag{119}$$

for some positive constants C_1, where the first inequality is by Minkowski's integral inequality, the second inequality is by the finiteness of $P_{\mathbf{X}}$, the third inequality is by Jensen's inequality, and the last equality is because of Parseval's identity.

Combining (118) and (119), we have

$$\begin{aligned}
\|f^* - f_n^*\|_{L_2(P_{\mathbf{X}})}^2 + \lambda_n\|f_n^*\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)}^2 \leq& C_1\|f^* - \tilde{f}_n^*\|_{L_2(\mathbb{R}^D)}^2 + \lambda_n\|\tilde{f}_n^*\|_{\mathcal{H}_{\tilde{K}_S}(\mathbb{R}^D)}^2\\
\leq& \max(C_1, 1)\left(\|f^* - \tilde{f}_n^*\|_{L_2(\mathbb{R}^D)}^2 + \lambda_n\|\tilde{f}_n^*\|_{\mathcal{H}_{\tilde{K}_S}(\mathbb{R}^D)}^2\right).
\end{aligned} \tag{120}$$

It remains to bound

$$\|f^* - \tilde{f}_n^*\|_{L_2(\mathbb{R}^D)}^2 + \lambda_n\|\tilde{f}_n^*\|_{\mathcal{H}_{\tilde{K}_S}(\mathbb{R}^D)}^2.$$

The Fourier inversion theorem implies that

$$\begin{aligned}
&\|f^* - \tilde{f}_n^*\|_{L_2(\mathbb{R}^D)}^2 + \lambda_n\|\tilde{f}_n^*\|_{\mathcal{H}_{\tilde{K}_S}(\mathbb{R}^D)}^2 = \int_{\mathbb{R}^D}|\mathcal{F}(f^*)(\boldsymbol{\omega}) - \mathcal{F}(\tilde{f}_n^*)(\boldsymbol{\omega})|^2 + \lambda_n\frac{|\mathcal{F}(\tilde{f}_n^*)(\boldsymbol{\omega})|^2}{\mathcal{F}(\tilde{K}_S)(\boldsymbol{\omega})}\mathrm{d}\boldsymbol{\omega}\\
\leq& \int_{\mathbb{R}^D}|\mathcal{F}(f^*)(\boldsymbol{\omega}) - \mathcal{F}(\tilde{g}_n^*)(\boldsymbol{\omega})|^2 + \lambda_n\frac{|\mathcal{F}(\tilde{g}_n^*)(\boldsymbol{\omega})|^2}{\mathcal{F}(\tilde{K}_S)(\boldsymbol{\omega})}\mathrm{d}\boldsymbol{\omega}\\
\leq& \int_{\mathbb{R}^D}|\mathcal{F}(f^*)(\boldsymbol{\omega}) - \mathcal{F}(\tilde{g}_n^*)(\boldsymbol{\omega})|^2 + C_2\lambda_n|\mathcal{F}(\tilde{g}_n^*)(\boldsymbol{\omega})|^2(1+\|\boldsymbol{\omega}\|_2^2)^{m_0}(1+\sigma_n^2\|\boldsymbol{\omega}\|_2^2)^{m_\varepsilon}\mathrm{d}\boldsymbol{\omega}\\
=& \int_{\mathbb{R}^D}\frac{C_2\lambda_n(1+\|\boldsymbol{\omega}\|_2^2)^{m_0}(1+\sigma_n^2\|\boldsymbol{\omega}\|_2^2)^{m_\varepsilon}}{1+C_2\lambda_n(1+\|\boldsymbol{\omega}\|_2^2)^{m_0}(1+\sigma_n^2\|\boldsymbol{\omega}\|_2^2)^{m_\varepsilon}}|\mathcal{F}(f^*)(\boldsymbol{\omega})|^2\mathrm{d}\boldsymbol{\omega}\\
\leq& \int_{\Omega_1}C_2\lambda_n(1+\|\boldsymbol{\omega}\|_2^2)^{m_0}(1+\sigma_n^2\|\boldsymbol{\omega}\|_2^2)^{m_\varepsilon}|\mathcal{F}(f^*)(\boldsymbol{\omega})|^2\mathrm{d}\boldsymbol{\omega}\\
&+\int_{\Omega_2}C_2\lambda_n(1+\|\boldsymbol{\omega}\|_2^2)^{m_0}(1+\sigma_n^2\|\boldsymbol{\omega}\|_2^2)^{m_\varepsilon}|\mathcal{F}(f^*)(\boldsymbol{\omega})|^2\mathrm{d}\boldsymbol{\omega} + \int_{\Omega_3}|\mathcal{F}(f^*)(\boldsymbol{\omega})|^2\mathrm{d}\boldsymbol{\omega}\\
=& I_1 + I_2 + I_3,
\end{aligned} \tag{121}$$

for some positive constants C_2, where $\tilde{g}_n^*$ minimizes

$$\int_{\mathbb{R}^D} |\mathcal{F}(f^*)(\boldsymbol{\omega}) - \mathcal{F}(\tilde{g}_n^*)(\boldsymbol{\omega})|^2 + C_2\lambda_n |\mathcal{F}(\tilde{g}_n^*)(\boldsymbol{\omega})|^2 (1+\|\boldsymbol{\omega}\|_2^2)^{m_0}(1+\sigma_n^2\|\boldsymbol{\omega}\|_2^2)^{m_\varepsilon} \mathrm{d}\boldsymbol{\omega},$$

$\Omega_1 = \{\boldsymbol{\omega} : C_2\lambda_n(1+\|\boldsymbol{\omega}\|_2^2)^{m_0}(1+\sigma_n^2\|\boldsymbol{\omega}\|_2^2)^{m_\varepsilon} \leq 1, \sigma_n^2\|\boldsymbol{\omega}\|_2^2 \leq m_\varepsilon^{-1}\}$, $\Omega_2 = \{\boldsymbol{\omega} : C_2\lambda_n(1+\|\boldsymbol{\omega}\|_2^2)^{m_0}(1+\sigma_n^2\|\boldsymbol{\omega}\|_2^2)^{m_\varepsilon} \leq 1, \sigma_n^2\|\boldsymbol{\omega}\|_2^2 \geq m_\varepsilon^{-1}\}$, and $\Omega_3 = \{\boldsymbol{\omega} : C_2\lambda_n(1+\|\boldsymbol{\omega}\|_2^2)^{m_0}(1+\sigma_n^2\|\boldsymbol{\omega}\|_2^2)^{m_\varepsilon} > 1\}$. In (121), the first inequality is because $\tilde{f}_n^*$ is the solution to the optimization problem (117), and the second inequality is by Assumption 4 (C1).

Since $\sigma_n^2\|\boldsymbol{\omega}\|_2^2 \leq m_\varepsilon^{-1}$ for $\boldsymbol{\omega} \in \Omega_1$, the first term I_1 in (121) can be bounded by

$$\begin{aligned} I_1 \leq & \int_{\Omega_1} C_2\lambda_n(1+\|\boldsymbol{\omega}\|_2^2)^{m_0}(1+m_\varepsilon^{-1})^{m_\varepsilon}|\mathcal{F}(f^*)(\boldsymbol{\omega})|^2 \mathrm{d}\boldsymbol{\omega} \\ \leq & C_2 e \int_{\Omega_1} \lambda_n (1+\|\boldsymbol{\omega}\|_2^2)^{m_0}|\mathcal{F}(f^*)(\boldsymbol{\omega})|^2 \mathrm{d}\boldsymbol{\omega}. \end{aligned} \tag{122}$$

If $m_0 \leq m_f$, then we directly have

$$I_1 \leq C_2 e \lambda_n \int_{\Omega_1} (1+\|\boldsymbol{\omega}\|_2^2)^{m_f}|\mathcal{F}(f^*)(\boldsymbol{\omega})|^2 d\boldsymbol{\omega}. \tag{123}$$

If $m_0 > m_f$, then for $\boldsymbol{\omega} \in \Omega_1$, we have

$$C_2\lambda_n(1+\|\boldsymbol{\omega}\|_2^2)^{m_0} \leq C_2\lambda_n(1+\|\boldsymbol{\omega}\|_2^2)^{m_0}(1+\sigma_n^2\|\boldsymbol{\omega}\|_2^2)^{m_\varepsilon} \leq 1,$$

which implies

$$C_2\lambda_n(1+\|\boldsymbol{\omega}\|_2^2)^{m_0} \leq \left(C_2\lambda_n(1+\|\boldsymbol{\omega}\|_2^2)^{m_0}\right)^{\frac{m_f}{m_0}} = C_3\lambda_n^{\frac{m_f}{m_0}}(1+\|\boldsymbol{\omega}\|_2^2)^{m_f},$$

for some positive constants C_3, therefore, by (122), we have

$$I_1 \leq C_4 e \lambda_n^{\frac{m_f}{m_0}} \int_{\Omega_1} (1+\|\boldsymbol{\omega}\|_2^2)^{m_f}|\mathcal{F}(f^*)(\boldsymbol{\omega})|^2 d\boldsymbol{\omega}. \tag{124}$$

for some positive constants C_4, The second term I_2 in (121) can be bounded by

$$\begin{aligned} I_2 \leq & \int_{\Omega_2} \left(C_2\lambda_n(1+\|\boldsymbol{\omega}\|_2^2)^{m_0}(1+\sigma_n^2\|\boldsymbol{\omega}\|_2^2)^{m_\varepsilon}\right)^{\frac{m_f}{m_0+m_\varepsilon}} |\mathcal{F}(f^*)(\boldsymbol{\omega})|^2 \mathrm{d}\boldsymbol{\omega} \\ \leq & \int_{\Omega_2} \left(C_2\lambda_n(1+\|\boldsymbol{\omega}\|_2^2)^{m_0}(m_\varepsilon+1)^{m_\varepsilon}\sigma_n^{2m_\varepsilon}\|\boldsymbol{\omega}\|_2^{2m_\varepsilon}\right)^{\frac{m_f}{m_0+m_\varepsilon}} |\mathcal{F}(f^*)(\boldsymbol{\omega})|^2 \mathrm{d}\boldsymbol{\omega} \\ \leq & \int_{\Omega_2} \left(C_2\lambda_n(m_\varepsilon+1)^{m_\varepsilon}\sigma_n^{2m_\varepsilon}(1+\|\boldsymbol{\omega}\|_2^2)^{m_0}(1+\|\boldsymbol{\omega}\|_2^2)^{m_\varepsilon}\right)^{\frac{m_f}{m_0+m_\varepsilon}} |\mathcal{F}(f^*)(\boldsymbol{\omega})|^2 \mathrm{d}\boldsymbol{\omega} \\ \leq & (C_2\lambda_n(m_\varepsilon+1)^{m_\varepsilon}\sigma_n^{2m_\varepsilon})^{\frac{m_f}{m_0+m_\varepsilon}} \int_{\Omega_2} (1+\|\boldsymbol{\omega}\|_2^2)^{m_f}|\mathcal{F}(f^*)(\boldsymbol{\omega})|^2 d\boldsymbol{\omega}, \end{aligned} \tag{125}$$

where the first inequality is because on Ω_2,

$$C_2\lambda_n(1+\|\boldsymbol{\omega}\|_2^2)^{m_0}(1+\sigma_n^2\|\boldsymbol{\omega}\|_2^2)^{m_\varepsilon} \leq 1,$$

implies

$$C_2\lambda_n(1+\|\boldsymbol{\omega}\|_2^2)^{m_0}(1+\sigma_n^2\|\boldsymbol{\omega}\|_2^2)^{m_\varepsilon} \leq \left(C_2\lambda_n(1+\|\boldsymbol{\omega}\|_2^2)^{m_0}(1+\sigma_n^2\|\boldsymbol{\omega}\|_2^2)^{m_\varepsilon}\right)^{\frac{m_f}{m_0+m_\varepsilon}},$$

provided $m_0 + m_\varepsilon \geq m_f$.

The third term I_3 in (121) can be bounded by

$$\begin{aligned} I_3 \leq & \int_{\Omega_3} \left(C_2\lambda_n(1+\|\boldsymbol{\omega}\|_2^2)^{m_0}(1+\sigma_n^2\|\boldsymbol{\omega}\|_2^2)^{m_\varepsilon}\right)^{\frac{m_f}{m_0+m_\varepsilon}} |\mathcal{F}(f^*)(\boldsymbol{\omega})|^2 \mathrm{d}\boldsymbol{\omega} \\ \leq & (C_2\lambda_n(m_\varepsilon+1)^{m_\varepsilon}\sigma_n^{2m_\varepsilon})^{\frac{m_f}{m_0+m_\varepsilon}} \int_{\Omega_3} (1+\|\boldsymbol{\omega}\|_2^2)^{m_f}|\mathcal{F}(f^*)(\boldsymbol{\omega})|^2 d\boldsymbol{\omega}, \end{aligned} \tag{126}$$

where the first inequality is because on Ω_3,

$$1 \leq C_2\lambda_n(1+\|\boldsymbol{\omega}\|_2^2)^{m_0}(1+\sigma_n^2\|\boldsymbol{\omega}\|_2^2)^{m_\varepsilon},$$

implies

$$1 \leq \left(C_2\lambda_n(1+\|\boldsymbol{\omega}\|_2^2)^{m_0}(1+\sigma_n^2\|\boldsymbol{\omega}\|_2^2)^{m_\varepsilon}\right)^{\frac{m_f}{m_0+m_\varepsilon}}.$$

Note that all constants C_j, $j = 1, ..., 4$ are not depending on m_ε. Furthermore, we have

$$C_2^{\frac{m_f}{m_0+m_\varepsilon}} \leq (\max(C_2,1))^{\frac{m_f}{m_0+m_\varepsilon}} \leq \max(C_2,1), \tag{127}$$

since $m_0 + m_\varepsilon \geq m_f$. By (127), plugging (123) (if $m_0 \leq m_f$) or (124) (if $m_0 > m_f$), (125), and (126) into (121), together with (120), finishes the proof. ■

I.2 Proof of Lemma 22

For $\boldsymbol{x} \in \Omega$, the Fourier inversion theorem implies

$$\begin{aligned} \tilde{K}_S(\boldsymbol{x}) = & \int_{\mathbb{R}^D}\int_{\mathbb{R}^D} K(\boldsymbol{x}+\boldsymbol{\varepsilon}-\boldsymbol{\varepsilon}')p_\varepsilon(\boldsymbol{\varepsilon})p_\varepsilon(\boldsymbol{\varepsilon}')\mathrm{d}\boldsymbol{\varepsilon}\mathrm{d}\boldsymbol{\varepsilon}' \\ = & (2\pi)^{-D/2}\int_{\mathbb{R}^D}\int_{\mathbb{R}^D}\int_{\mathbb{R}^D} e^{-i(\boldsymbol{x}+\boldsymbol{\varepsilon}-\boldsymbol{\varepsilon}')^T\boldsymbol{\omega}}\mathcal{F}(K)(\boldsymbol{\omega})\mathrm{d}\boldsymbol{\omega}p_\varepsilon(\boldsymbol{\varepsilon})p_\varepsilon(\boldsymbol{\varepsilon}')\mathrm{d}\boldsymbol{\varepsilon}\mathrm{d}\boldsymbol{\varepsilon}' \\ = & (2\pi)^{-D/2}\int_{\mathbb{R}^D} e^{-i\boldsymbol{x}^T\boldsymbol{\omega}}\mathcal{F}(K)(\boldsymbol{\omega})|\varphi_\varepsilon(\boldsymbol{\omega})|^2\mathrm{d}\boldsymbol{\omega}, \end{aligned}$$

where φ_ε is the characteristic function of p_ε. Thus, by the Fourier theorem,

$$\mathcal{F}(\tilde{K}_S(\boldsymbol{x}))(\boldsymbol{\omega}) = \mathcal{F}(K)(\boldsymbol{\omega})|\varphi_\varepsilon(\boldsymbol{\omega})|^2. \tag{128}$$

Let Ψ_σ be a positive definite function satisfying

$$c_1\left(1+\frac{\sigma^2}{m_0+m_\varepsilon}\|\boldsymbol{\omega}\|_2^2\right)^{-(m_0+m_\varepsilon)} \leq \mathcal{F}(\Psi_\sigma) \leq c_2\left(1+\frac{\sigma^2}{m_0+m_\varepsilon}\|\boldsymbol{\omega}\|_2^2\right)^{-(m_0+m_\varepsilon)}, \forall \boldsymbol{\omega} \in \mathbb{R}^D,$$

and $\mathcal{N}_\sigma(\Omega)$ be the RKHS generated by Ψ_σ, where the constants c_1 and c_2 are not depending on m_ε. Therefore, for any $f \in \mathcal{H}_{\tilde{K}_S}(\Omega)$, we have that

$$\begin{aligned}
\|f\|^2_{\mathcal{N}_{\sigma_n}(\Omega)} =& \int_{\mathbb{R}^D} \frac{|\mathcal{F}(f)(\boldsymbol{\omega})|^2}{\mathcal{F}(\Psi_\sigma)(\boldsymbol{\omega})} \mathrm{d}\boldsymbol{\omega} \\
\leq & C_1 \int_{\mathbb{R}^D} \left(1 + \frac{\sigma^2}{m_0 + m_\varepsilon}\|\boldsymbol{\omega}\|_2^2\right)^{m_0+m_\varepsilon} |\mathcal{F}(f)(\boldsymbol{\omega})|^2 \mathrm{d}\boldsymbol{\omega} \\
\leq & C_1 \int_{\mathbb{R}^D} (1 + \|\boldsymbol{\omega}\|_2^2)^{m_0} (1 + \sigma^2\|\boldsymbol{\omega}\|_2^2)^{m_\varepsilon} |\mathcal{F}(f)(\boldsymbol{\omega})|^2 \mathrm{d}\boldsymbol{\omega} \\
\leq & C_2 \int_{\mathbb{R}^D} \frac{|\mathcal{F}(f)(\boldsymbol{\omega})|^2}{\mathcal{F}(K)(\boldsymbol{\omega})|\varphi_\varepsilon(\boldsymbol{\omega})|^2} \mathrm{d}\boldsymbol{\omega} \\
= & C_2 \int_{\mathbb{R}^D} \frac{|\mathcal{F}(f)(\boldsymbol{\omega})|^2}{\mathcal{F}(\tilde{K}_S)(\boldsymbol{\omega})} \mathrm{d}\boldsymbol{\omega},
\end{aligned}$$

for some positive constants C_1 and C_2, provided $\sigma \leq 1$, where the last inequality is because of Assumptions 2 and 4 (C1). Thus, we have if $\sigma \leq 1$,

$$\|f\|_{\mathcal{N}_\sigma(\Omega)} \leq C_3 \|f\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)}, \tag{129}$$

for some positive constants C_3.

In order to prove Lemma 22, we need the following lemmas. Although we can directly apply Corollary A.8 of Hamm and Steinwart (2021a) and the entropy number of Sobolev spaces to obtain an upper bound on $H(\delta, \mathcal{B}_{\mathcal{H}_\sigma(\Omega)}, \|\cdot\|_{L_\infty(\Omega)})$, which is

$$H(\delta, \mathcal{B}_{\mathcal{H}_\sigma(\Omega)}, \|\cdot\|_{L_\infty(\Omega)}) \leq C\sigma^{-d}\delta^{-\frac{D}{m_0+m_\varepsilon}}, \tag{130}$$

where C is a constant *depending on* m_ε. However, the dependency between C and m_ε is not clear as far as we know, and thus cannot meet our needs when m_ε is dependent on the sample size n. Therefore, we develop Lemma 39, providing a new upper bound on $H(\delta, \mathcal{B}_{\mathcal{H}_m([0,1]^D)}, \|\cdot\|_{L_\infty([0,1]^D)})$, where the dependency between the upper bound and m_ε is clearly described. Based on Lemma 39, we provide Lemma 32, where the constant is independent with m_ε.

Lemma 33 is a Bernstein-type inequality for a single g. See, for example, Massart (2007).

Lemma 32 *Suppose the conditions of Lemma 22 are fulfilled. Let $\mathcal{B}_{\mathcal{N}_\sigma(\Omega)}$ be a unit ball in $\mathcal{N}_\sigma(\Omega)$. Then for all $\delta > 0$, we have*

$$H(\delta, \mathcal{B}_{\mathcal{H}_\sigma(\Omega)}, \|\cdot\|_{L_\infty(\Omega)}) \leq C\sigma^{-d}(2m - D)^{-\frac{2D}{2m-D}} m^{\frac{2mD}{2m-D}} \delta^{-\frac{2D}{2m-D}} \log(1 + \delta^{-1}),$$

where the constant C is independent with m_ε, and $m = m_\varepsilon + m_0$.

Lemma 33 *Suppose $X_i \sim Unif(\Omega)$ for $i = 1, \ldots, n$. Let g be a fixed function. We have for all $t > 0$,*

$$P\left(\left|\|g\|_n^2 - \|g\|^2_{L_2(P_{\mathbf{X}})}\right| \geq t\right) \leq 2\exp\left(-\frac{nt^2}{8(t + \|g\|^2_{L_2(P_{\mathbf{X}})})}\right).$$

Lemma 34 *Suppose conditions of Theorem 9 are fulfilled. Then for some constant $C_2 > 0$ only related to Assumption 1 and for $\delta > 0$ with*

$$\sqrt{n}\delta > 2C_2 \max\left(\int_0^1 H(u, \mathcal{B}_{\mathcal{H}_\sigma(\Omega)}, \|\cdot\|_{L_\infty(\Omega)})^{1/2}\mathrm{d}u, 1\right),$$

we have for $p = \frac{4D}{2(m_0+m_\varepsilon)-D}$, $m = m_0 + m_\varepsilon$, and $\sqrt{n}\delta \geq C\sigma^{-d/2}m^{\frac{mD}{2m-D}+\frac{1}{2}}$,

$$\mathbb{P}\left(\sup_{g\in\mathcal{B}_{\mathcal{H}_\sigma(\Omega)}} \frac{\langle g, \boldsymbol{\epsilon}\rangle_n}{\|g\|_n^{1-\frac{p}{2}}} \geq \delta\right) \leq C_3 p^{-1}\exp\left(-\frac{n\delta^2}{C_3^2}\right),$$

where the constants C, C_2 and C_3 are independent with m_ε.

Proof of Lemma 34. The proof can be obtained by applying the peeling-off argument in Lemma 8.4 of van de Geer (2000). Let $m = m_0 + m_\varepsilon$. Note that

$$\begin{aligned}
&\int_0^\delta H(u, \mathcal{B}_{\mathcal{H}_\sigma(\Omega)}, \|\cdot\|_{L_\infty(\Omega)})^{1/2}\mathrm{d}u \\
\leq & C\sigma^{-d/2}(2m-D)^{-\frac{D}{2m-D}}m^{\frac{mD}{2m-D}}\int_0^\delta u^{-\frac{D}{2m-D}}\sqrt{\log(1+u^{-1})}\mathrm{d}u \\
\leq & C\sigma^{-d/2}(2m-D)^{-\frac{D}{2m-D}}m^{\frac{mD}{2m-D}}\int_0^\delta u^{-\frac{D}{2m-D}}\sqrt{\frac{2m-D}{2D}\left(1+\frac{1}{u}\right)^{\frac{2D}{2m-D}}}\mathrm{d}u \\
\leq & C_1\sigma^{-d/2}(2m-D)^{-\frac{D}{2m-D}}m^{\frac{mD}{2m-D}+\frac{1}{2}}\int_0^\delta u^{-\frac{2D}{2m-D}}\mathrm{d}u \\
= & C_1\sigma^{-d/2}(2m-D)^{-\frac{D}{2m-D}}m^{\frac{mD}{2m-D}+\frac{1}{2}}\left(1-\frac{2D}{2m-D}\right)^{-1}\delta^{1-\frac{2D}{2m-D}} \\
\leq & C_1\sigma^{-d/2}(2m_f+2D)^{-\frac{D}{2m_f+2D}}m^{\frac{mD}{2m-D}+\frac{1}{2}}\left(1-\frac{D}{m_f+1}\right)^{-1}\delta^{1-\frac{2D}{2m-D}} \\
= & C_2\sigma^{-d/2}m^{\frac{mD}{2m-D}+\frac{1}{2}}\delta^{1-\frac{2D}{2m-D}},
\end{aligned}$$

for some positive constants C and C_1, and C_2, where the first inequality is by Lemma 32, the second inequality is by the basic inequality $\log(1+1/u) \leq a(1+1/u)^{1/a}$ for any $u, a > 0$, and the fourth inequality holds as long as $m_\varepsilon \geq m_f + D$. Here the constant C_2 is independent of m.

Let $p = \frac{4D}{2m-D}$ and $\sqrt{n}\delta \geq 4CC_2\sigma^{-d/2}m^{\frac{mD}{2m-D}+\frac{1}{2}}$, where C is only depending on Assumption 1. The proof then follows the proof of Lemma 8.4 of van de Geer (2000), while the last step becomes

$$\mathbb{P}\left(\sup_{g\in\mathcal{B}_{\mathcal{H}_\sigma(\Omega)}} \frac{\langle g, \boldsymbol{\epsilon}\rangle_n}{\|g\|_n^{1-\frac{p}{2}}} \geq \delta\right) \leq \sum_{s=1}^\infty C_3\exp\left(-\frac{n\delta^2}{16C_3^2}2^{sp}\right) \leq C_4 p^{-1}\exp\left(-\frac{n\delta^2}{C_4^2}\right),$$

for some positive constants C_3 and C_4, where we use a similar approach in the proof of Lemma 36. ■

Proof of Lemma 22. Since $\hat{f}$ is the solution to the optimization problem (39), it can be seen that

$$\|\hat{f}_n - \boldsymbol{y}\|_n^2 + \lambda_n \|\hat{f}_n\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)} \leq \|f_n^* - \boldsymbol{y}\|_n^2 + \lambda_n \|f_n^*\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)}, \tag{131}$$

where f_n^* is as in Lemma 21. By rearrangement, (131) implies

$$\|f^* - \hat{f}_n\|_n^2 + C_5\lambda_n \|\hat{f}_n\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)} \leq \|f^* - f_n^*\|_n^2 + C_6\lambda_n \|f_n^*\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)} + 2\langle \boldsymbol{\epsilon}, \hat{f}_n - f_n^*\rangle_n \tag{132}$$

for some positive constants C_5 and C_6. Take

$$\delta_n = 4CC_2 n^{-1/2}\sigma_n^{-d/2} m^{\frac{mD}{2m-D}+\frac{1}{2}},$$

and let $p = \frac{4D}{2m-D}$, where $m = m_0 + m_\varepsilon$. Applying Lemma 34, with probability at least

$$C_6 p^{-1}\exp\left(-C_7\sigma^{-d} m^{\frac{2mD}{2m-D}+1}\right),$$

for some positive constants C_7, which converges to zero by our assumption, we have

$$2\langle \boldsymbol{\epsilon}, \hat{f}_n - f_n^*\rangle_n \leq C_8 n^{-1/2}\sigma^{-d/2} m^{\frac{mD}{2m-D}+\frac{1}{2}} \|\hat{f}_n - f_n^*\|_n^{1-\frac{p}{2}} (\|\hat{f}_n\|_{\mathcal{N}_{\sigma_n}(\Omega)} + \|f_n^*\|_{\mathcal{N}_{\sigma_n}(\Omega)})^{\frac{p}{2}}$$

for some positive constants C_8, which, together with (132), implies

$$\begin{aligned}
&\|f^* - \hat{f}_n\|_n^2 + \lambda_n \|\hat{f}_n\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)} \\
\leq &\|f^* - f_n^*\|_n^2 + \lambda_n \|f_n^*\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)} \\
&+ C_8 n^{-1/2}\sigma^{-d/2} m^{\frac{mD}{2m-D}+\frac{1}{2}} \|\hat{f}_n - f_n^*\|_n^{1-\frac{p}{2}} (\|\hat{f}_n\|_{\mathcal{N}_{\sigma_n}(\Omega)} + \|f_n^*\|_{\mathcal{N}_{\sigma_n}(\Omega)})^{\frac{p}{2}}.
\end{aligned} \tag{133}$$

By assumption of Lemma 22, we have

$$\|f^* - f_n^*\|^2_{L_2(P_{\mathbf{X}})} + \lambda_n \|f_n^*\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)} \leq T, \tag{134}$$

which implies $\|f_n^*\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)} = O(\lambda_n^{-1}T)$.

Now we consider bounding the difference between $\|f^* - f_n^*\|_n$ and $\|f^* - f_n^*\|_{L_2(P_{\mathbf{X}})}$. Since f_n^* does not depend on $\boldsymbol{x}_j$'s and $\boldsymbol{\epsilon}$, we can directly apply Lemma 33 to $\|f^* - f_n^*\|_n$ and obtain that

$$\left| \|f^* - f_n^*\|_n^2 - \|f^* - f_n^*\|^2_{L_2(P_{\mathbf{X}})} \right| = O_{\mathbb{P}}(n^{-1/2})\|f^* - f_n^*\|_{L_2(P_{\mathbf{X}})},$$

which, together with (134), yields

$$\|f^* - f_n^*\|_n^2 = O_{\mathbb{P}}\left(T + n^{-1/2}T^{1/2}\right). \tag{135}$$

Plugging (135) into (133), together with (134), gives us

$$
\begin{aligned}
&\|f^* - \hat{f}_n\|_n^2 + \lambda_n \|\hat{f}_n\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)} \\
=&O_{\mathbb{P}}\left(T + n^{-1/2}T^{1/2}\right) \\
&+ O_{\mathbb{P}}\left(n^{-1/2}\sigma^{-d/2}m^{\frac{mD}{2m-D}+\frac{1}{2}}\|\hat{f}_n - f_n^*\|_n^{1-\frac{p}{2}}(\|\hat{f}_n\|_{\mathcal{N}_{\sigma_n}(\Omega)} + \|f_n^*\|_{\mathcal{N}_{\sigma_n}(\Omega)})^{\frac{p}{2}}\right), \qquad (136)
\end{aligned}
$$

where we also use $\|f\|_{\mathcal{N}_{\sigma_n}(\Omega)} \le C_3 \|f\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)}$ for all $f \in \mathcal{H}_{\tilde{K}_S}(\Omega)$ (see (129)). Then (136) implies either

$$
\|f^* - \hat{f}_n\|_n^2 + \lambda_n \|\hat{f}_n\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)} = O_{\mathbb{P}}\left(T + n^{-1/2}T^{1/2}\right), \qquad (137)
$$

or

$$
\begin{aligned}
&\|f^* - \hat{f}_n\|_n^2 + \lambda_n \|\hat{f}_n\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)} \\
=&O_{\mathbb{P}}\left(n^{-1/2}\sigma^{-d/2}m^{\frac{mD}{2m-D}+\frac{1}{2}}\|\hat{f}_n - f_n^*\|_n^{1-\frac{p}{2}}(\|\hat{f}_n\|_{\mathcal{N}_{\sigma_n}(\Omega)} + \|f_n^*\|_{\mathcal{N}_{\sigma_n}(\Omega)})^{\frac{p}{2}}\right). \qquad (138)
\end{aligned}
$$

In order to solve (138), we consider two cases.

Case 1: $\|\hat{f}_n\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)} \ge \|f_n^*\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)}$. In this case, we have

$$
\begin{aligned}
&\|f^* - \hat{f}_n\|_n^2 + \lambda_n \|\hat{f}_n\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)} = O_{\mathbb{P}}\left(\sigma_n^{-d/2}n^{-1/2}m^{\frac{mD}{2m-D}+\frac{1}{2}}\|\hat{f}_n - f_n^*\|_n^{1-\frac{p}{2}}\|\hat{f}_n\|^{\frac{p}{2}}_{\mathcal{H}_{\tilde{K}_S}(\Omega)}\right) \\
=&O_{\mathbb{P}}\left(\sigma_n^{-d/2}n^{-1/2}m^{\frac{mD}{2m-D}+\frac{1}{2}}\|f^* - f_n^*\|_n^{1-\frac{p}{2}}\|\hat{f}_n\|^{\frac{p}{2}}_{\mathcal{H}_{\tilde{K}_S}(\Omega)}\right) \\
&+ O_{\mathbb{P}}\left(\sigma_n^{-d/2}n^{-1/2}m^{\frac{mD}{2m-D}+\frac{1}{2}}\|f^* - \hat{f}_n\|_n^{1-\frac{p}{2}}\|\hat{f}_n\|^{\frac{p}{2}}_{\mathcal{H}_{\tilde{K}_S}(\Omega)}\right), \qquad (139)
\end{aligned}
$$

where the second equality (with $O_{\mathbb{P}}$ notation) is because of the triangle inequality and the basic inequality $(a+b)^q \le a^q + b^q$ for $q \in (0,1)$.

It can be seen that (139) further implies

$$
\|f^* - \hat{f}_n\|_n^2 + \lambda_n \|\hat{f}_n\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)} = O_{\mathbb{P}}\left(\sigma_n^{-d/2}n^{-1/2}m^{\frac{mD}{2m-D}+\frac{1}{2}}\|f^* - f_n^*\|_n^{1-\frac{p}{2}}\|\hat{f}_n\|^{\frac{p}{2}}_{\mathcal{H}_{\tilde{K}_S}(\Omega)}\right), \qquad (140)
$$

or

$$
\|f^* - \hat{f}_n\|_n^2 + \lambda_n \|\hat{f}_n\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)} = O_{\mathbb{P}}\left(\sigma_n^{-d/2}n^{-1/2}m^{\frac{mD}{2m-D}+\frac{1}{2}}\|f^* - \hat{f}_n\|_n^{1-\frac{p}{2}}\|\hat{f}_n\|^{\frac{p}{2}}_{\mathcal{H}_{\tilde{K}_S}(\Omega)}\right). \qquad (141)
$$

Plugging (135) into (140), we have

$$
\|f^* - \hat{f}_n\|_n^2 + \lambda_n \|\hat{f}_n\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)} = O_{\mathbb{P}}\left(\sigma_n^{-d/2}n^{-1/2}m^{\frac{mD}{2m-D}+\frac{1}{2}}(T + n^{-1/2}T^{1/2})^{\frac{1}{2}-\frac{p}{4}}\|\hat{f}_n\|^{\frac{p}{2}}_{\mathcal{H}_{\tilde{K}_S}(\Omega)}\right). \qquad (142)
$$

Solving (142) yields

$$\|f^* - \hat{f}_n\|_n = O_{\mathbb{P}}\left(\lambda_n^{-\frac{p}{2(4-p)}}\left(\sigma_n^{-d/2}n^{-1/2}m^{\frac{mD}{2m-D}+\frac{1}{2}}(T+n^{-1/2}T^{1/2})^{\frac{1}{2}-\frac{p}{4}}\right)^{\frac{2}{4-p}}\right),$$
$$\|\hat{f}_n\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)} = O_{\mathbb{P}}\left(\left(\lambda_n^{-1}\sigma_n^{-d/2}m^{\frac{mD}{2m-D}+\frac{1}{2}}n^{-1/2}(T+n^{-1/2}T^{1/2})^{\frac{1}{2}-\frac{p}{4}}\right)^{\frac{2}{4-p}}\right). \quad (143)$$

Solving (141) yields

$$\|f^* - \hat{f}_n\|_n = O_{\mathbb{P}}\left(\sigma_n^{-d/2}n^{-1/2}m^{\frac{mD}{2m-D}+\frac{1}{2}}\lambda_n^{-\frac{p}{4}}\right),$$
$$\|\hat{f}_n\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)} = O_{\mathbb{P}}\left(\sigma_n^{-d/2}n^{-1/2}m^{\frac{mD}{2m-D}+\frac{1}{2}}\lambda_n^{-\frac{2+p}{4}}\right). \quad (144)$$

Case 2: $\|\hat{f}_n\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)} < \|f_n^*\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)}$. In this case, (138) implies that

$$\|f^* - \hat{f}_n\|_n^2 + \lambda_n\|\hat{f}_n\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)} = O_{\mathbb{P}}\left(\sigma_n^{-d/2}n^{-1/2}m^{\frac{mD}{2m-D}+\frac{1}{2}}\|\hat{f}_n - f_n^*\|_n^{1-\frac{p}{2}}\|f_n^*\|^{\frac{p}{2}}_{\mathcal{H}_{\tilde{K}_S}(\Omega)}\right)$$
$$= O_{\mathbb{P}}\left(\sigma_n^{-d/2}n^{-1/2}m^{\frac{mD}{2m-D}+\frac{1}{2}}\|f^* - f_n^*\|_n^{1-\frac{p}{2}}\|f_n^*\|^{\frac{p}{2}}_{\mathcal{H}_{\tilde{K}_S}(\Omega)}\right)$$
$$+ O_{\mathbb{P}}\left(\sigma_n^{-d/2}n^{-1/2}m^{\frac{mD}{2m-D}+\frac{1}{2}}\|f^* - \hat{f}_n\|_n^{1-\frac{p}{2}}\|f_n^*\|^{\frac{p}{2}}_{\mathcal{H}_{\tilde{K}_S}(\Omega)}\right), \quad (145)$$

where the second equality is because of the triangle inequality and the basic inequality $(a+b)^q \le a^q + b^q$ for $q \in (0,1)$ again.

By (145), we have either

$$\|f^* - \hat{f}_n\|_n^2 + \lambda_n\|\hat{f}_n\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)} = O_{\mathbb{P}}\left(\sigma_n^{-d/2}n^{-1/2}m^{\frac{mD}{2m-D}+\frac{1}{2}}\|f^* - f_n^*\|_n^{1-\frac{p}{2}}\|f_n^*\|^{\frac{p}{2}}_{\mathcal{H}_{\tilde{K}_S}(\Omega)}\right), \quad (146)$$

or

$$\|f^* - \hat{f}_n\|_n^2 + \lambda_n\|\hat{f}_n\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)} = O_{\mathbb{P}}\left(\sigma_n^{-d/2}n^{-1/2}m^{\frac{mD}{2m-D}+\frac{1}{2}}\|f^* - \hat{f}_n\|_n^{1-\frac{p}{2}}\|f_n^*\|^{\frac{p}{2}}_{\mathcal{H}_{\tilde{K}_S}(\Omega)}\right). \quad (147)$$

Combining (146) and (135), we have

$$\|f^* - \hat{f}\|_n^2 = O_{\mathbb{P}}\left(\sigma_n^{-d/2}n^{-1/2}m^{\frac{mD}{2m-D}+\frac{1}{2}}(\lambda_n^{-1}T)^{\frac{p}{2}}(T+n^{-1/2}T^{1/2})^{1-\frac{p}{2}}\right),$$
$$\|\hat{f}\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)} = O_{\mathbb{P}}\left(\lambda_n^{-1}\sigma_n^{-d/2}n^{-1/2}m^{\frac{mD}{2m-D}+\frac{1}{2}}(\lambda_n^{-1}T)^{\frac{p}{2}}(T+n^{-1/2}T^{1/2})^{1-\frac{p}{2}}\right). \quad (148)$$

Combining (147) and (135), we have

$$\|f^* - \hat{f}_n\|_n = O_{\mathbb{P}}\left((\sigma_n^{-d/2}n^{-1/2}m^{\frac{mD}{2m-D}+\frac{1}{2}})^{\frac{2}{2+p}}(\lambda_n^{-1}T)^{\frac{p}{2(2+p)}}\right)$$
$$\|\hat{f}_n\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)} = O_{\mathbb{P}}\left(\lambda_n^{-1/2}(\sigma_n^{-d/2}n^{-1/2}m^{\frac{mD}{2m-D}+\frac{1}{2}})^{\frac{2}{2+p}}(\lambda_n^{-1}T)^{\frac{p}{2(2+p)}}\right). \quad (149)$$

By (137), (144), (143), (148), and (149), we finish the proof. ■

I.3 Proof of Lemma 24

For any function $g \in \mathcal{W}^m(\mathbb{R}^D)$ where $m = m_0 + m_\varepsilon$, the Fourier inversion theorem implies

$$
\begin{aligned}
|g(\boldsymbol{x})| &= \left| \int_{\mathbb{R}^D} e^{i\boldsymbol{x}^T\boldsymbol{\omega}} \mathcal{F}(g)(\boldsymbol{\omega}) d\boldsymbol{\omega} \right| \leq \int_{\mathbb{R}^D} |\mathcal{F}(g)(\boldsymbol{\omega})| \, d\boldsymbol{\omega} \\
&= \int_{\mathbb{R}^D} |\mathcal{F}(g)(\boldsymbol{\omega})|^{1-r} (\mathcal{F}(k_\sigma)(\boldsymbol{\omega}))^{r/2} |\mathcal{F}(g)(\boldsymbol{\omega})|^r (\mathcal{F}(k_\sigma)(\boldsymbol{\omega}))^{-r/2} d\boldsymbol{\omega} \\
&\leq \left(\int_{\mathbb{R}^D} |\mathcal{F}(g)(\boldsymbol{\omega})|^{\frac{2(1-r)}{2-r}} (\mathcal{F}(k_\sigma)(\boldsymbol{\omega}))^{\frac{r}{2-r}} d\boldsymbol{\omega} \right)^{\frac{2-r}{2}} \left(\int_{\mathbb{R}^D} |\mathcal{F}(g)(\boldsymbol{\omega})|^2 (\mathcal{F}(k_\sigma)(\boldsymbol{\omega}))^{-1} d\boldsymbol{\omega} \right)^{\frac{r}{2}} \\
&\leq \left(\int_{\mathbb{R}^D} |\mathcal{F}(g)(\boldsymbol{\omega})|^{\frac{2(1-r)}{2-r}} \left|(1 + \|\boldsymbol{\omega}\|_2^2)^{-m}\right|^{\frac{r}{2-r}} d\boldsymbol{\omega} \right)^{\frac{2-r}{2}} \|g\|_{\mathcal{H}_\sigma(\mathbb{R}^D)}^r \\
&\leq \left(\int_{\mathbb{R}^D} |\mathcal{F}(g)(\boldsymbol{\omega})|^2 \, d\boldsymbol{\omega} \right)^{\frac{1-r}{2}} \left(\int_{\mathbb{R}^D} (1 + \|\boldsymbol{\omega}\|_2^2)^{-mr} d\boldsymbol{\omega} \right)^{\frac{1}{2}} \|g\|_{\mathcal{H}_\sigma(\mathbb{R}^D)}^r \\
&= \left(\int_{\mathbb{R}^D} (1 + \|\boldsymbol{\omega}\|_2^2)^{-mr} d\boldsymbol{\omega} \right)^{\frac{1}{2}} \|g\|_{L_2(\mathbb{R}^D)}^{1-r} \|g\|_{\mathcal{H}_\sigma(\mathbb{R}^D)}^r,
\end{aligned} \tag{150}
$$

where the second and fourth inequalities are by Hölder's inequality, and the third equality is by Parseval's identity. Taking $r = \frac{D}{2(m_0+m_\varepsilon)}$ in (150), we have

$$
\begin{aligned}
|g(\boldsymbol{x})| &\leq \left(\int_{\mathbb{R}^D} (1 + \|\boldsymbol{\omega}\|_2^2)^{-mr} d\boldsymbol{\omega} \right)^{\frac{1}{2}} \|g\|_{L_2(\mathbb{R}^D)}^{1-r} \|g\|_{\mathcal{H}_\sigma(\mathbb{R}^D)}^r \\
&= \left(\int_{\mathbb{R}^D} (1 + \|\boldsymbol{\omega}\|_2^2)^{-\frac{D}{2}} d\boldsymbol{\omega} \right)^{\frac{1}{2}} \|g\|_{L_2(\mathbb{R}^D)}^{1-r} \|g\|_{\mathcal{H}_\sigma(\mathbb{R}^D)}^r \\
&= C_4 \|g\|_{L_2(\mathbb{R}^D)}^{1-r} \|g\|_{\mathcal{H}_\sigma(\mathbb{R}^D)}^r,
\end{aligned}
$$

for some positive constants C_4. This finishes the proof. ■

Appendix J. Proof of Lemmas in Appendix E

J.1 Proof of Lemma 25

By Theorem 10.46 of Wendland (2004), there exists a nature extension of $f \in \mathcal{H}_{\tilde{K}_S}(\Omega)$ on $\mathbb{R}^D$, such that the RKHS norm is preserved. Thus, we can focus on the RKHS $\mathcal{H}_{\tilde{K}_S}(\mathbb{R}^D)$.

By (128), we have that for any $f \in \mathcal{H}_{\tilde{K}_S}(\mathbb{R}^D)$,

$$
\|f\|_{\mathcal{H}_{\tilde{K}_S}(\mathbb{R}^D)}^2 = \int_{\mathbb{R}^D} \frac{|\mathcal{F}(f)(\boldsymbol{\omega})|^2}{\mathcal{F}(K)(\boldsymbol{\omega}) |\varphi_\varepsilon(\boldsymbol{\omega})|^2} \mathrm{d}\boldsymbol{\omega}.
$$

For normal distribution, the characteristic function satisfies $\varphi_\varepsilon(\boldsymbol{\omega}) = e^{-\frac{1}{2}\sigma_n^2 \|\boldsymbol{\omega}\|_2^2}$. Let $g_1(u) = \sigma_n^2 u - m_0 \log(1+u)$. Taking the derivative, we obtain

$$
g_1'(u) = \sigma_n^2 - \frac{m_0}{1+u},
$$

which is smaller than zero when $u \in [0, \frac{m_0}{\sigma_n^2} - 1)$, and larger than zero when $u \in (\frac{m_0}{\sigma_n^2} - 1, \infty)$. Therefore,

$$\begin{aligned} g_1(u) \geq & g_1\left(\frac{m_0}{\sigma_n^2} - 1\right) = m_0 - \sigma_n^2 - m_0 \log m_0 + 2m_0 \log \sigma_n \\ \geq & m_0 - 1 - m_0 \log m_0 + 2m_0 \log \sigma_n, \forall u \in [0, \infty), \end{aligned}$$

which implies

$$(1+u)^{-m_0} e^{\sigma_n^2 u} \geq e^{C_1} \sigma_n^{2m_0},$$

where $C_1 = m_0 - 1 - m_0 \log m_0$. By taking $u = \|\boldsymbol{\omega}\|_2^2$, Assumption 3 implies

$$\mathcal{F}(K)(\boldsymbol{\omega})|\varphi_\varepsilon(\boldsymbol{\omega})|^2 \geq c_1(1 + \|\boldsymbol{\omega}\|_2^2)^{-m_0} e^{-2\sigma_n^2 \|\boldsymbol{\omega}\|_2^2} \geq C_2 \sigma_n^{2m_0} e^{-3\sigma_n^2 \|\boldsymbol{\omega}\|_2^2}, \tag{151}$$

for some positive constants C_4. As for an upper bound of $\mathcal{F}(K)(\boldsymbol{\omega})|\varphi_\varepsilon(\boldsymbol{\omega})|^2$, direct computation shows that

$$\mathcal{F}(K)(\boldsymbol{\omega})|\varphi_\varepsilon(\boldsymbol{\omega})|^2 \leq c_2(1 + \|\boldsymbol{\omega}\|_2^2)^{-m_0} e^{-\sigma_n^2 \|\boldsymbol{\omega}\|_2^2} \leq c_2 e^{-\frac{1}{2}\sigma_n^2 \|\boldsymbol{\omega}\|_2^2}. \tag{152}$$

By (66), the Fourier transform of $k_\sigma(\cdot)$ is

$$\mathcal{F}(k_\sigma)(\boldsymbol{\omega}) = (2\sigma)^D e^{-\sigma^2 \|\boldsymbol{\omega}\|_2^2}. \tag{153}$$

Let $\mathcal{H}_\sigma(\mathbb{R}^D)$ be the RKHS generated by $k_\sigma(\boldsymbol{x} - \boldsymbol{x}')$. From (151), (152), and (153), it can be seen that

$$\begin{aligned} \|h_1\|^2_{\mathcal{H}_{\tilde{K}_S}(\mathbb{R}^D)} = & \int_{\mathbb{R}^D} \frac{|\mathcal{F}(f)(\boldsymbol{\omega})|^2}{\mathcal{F}(K)(\boldsymbol{\omega})|\varphi_\varepsilon(\boldsymbol{\omega})|^2} \mathrm{d}\boldsymbol{\omega} \\ \geq & C_3 \int_{\mathbb{R}^D} |\mathcal{F}(f)(\boldsymbol{\omega})|^2 e^{\frac{1}{2}\sigma_n^2 \|\boldsymbol{\omega}\|_2^2} \mathrm{d}\boldsymbol{\omega} \\ \geq & C_4 \sigma_n^D \|h_1\|^2_{\mathcal{H}_{\sigma_n/\sqrt{2}}(\mathbb{R}^D)}, \end{aligned}$$

and

$$\begin{aligned} \|h_2\|^2_{\mathcal{H}_{\tilde{K}_S}(\mathbb{R}^D)} = & \int_{\mathbb{R}^D} \frac{|\mathcal{F}(f)(\boldsymbol{\omega})|^2}{\mathcal{F}(K)(\boldsymbol{\omega})|\varphi_\varepsilon(\boldsymbol{\omega})|^2} \mathrm{d}\boldsymbol{\omega} \\ \leq & C_5 \int_{\mathbb{R}^D} |\mathcal{F}(f)(\boldsymbol{\omega})|^2 \sigma_n^{-2m_0} e^{3\sigma_n^2 \|\boldsymbol{\omega}\|_2^2} \mathrm{d}\boldsymbol{\omega} \\ \leq & C_6 \sigma_n^{-2m_0 - D} \|h_2\|^2_{\mathcal{H}_{\sqrt{3}\sigma_n}(\mathbb{R}^D)}, \end{aligned}$$

for some positive constants C_5, C_6, $h_1 \in \mathcal{H}_{\tilde{K}_S}(\mathbb{R}^D)$ and $h_2 \in \mathcal{H}_{\sqrt{3}\sigma_n}(\mathbb{R}^D)$, where C_4 and C_6 does not depend on σ_n. ■

J.2 Proof of Lemma 26

By (119), the Fourier inversion theorem, and Parseval's identity, it can be shown that

$$\begin{aligned}
&\|f^* - f_n^*\|_{L_2(P_{\mathbf{X}})}^2 + \lambda_n \|f_n^*\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)}^2 \\
\leq & C_1 \left(\|f^* - f_n^*\|_{L_2(\mathbb{R}^D)}^2 + \lambda_n \|f_n^*\|_{\mathcal{H}_{\tilde{K}_S}(\mathbb{R}^D)}^2 \right) \\
= & C_1 \left(\int_{\mathbb{R}^D} |\mathcal{F}(f^*)(\boldsymbol{\omega}) - \mathcal{F}(f_n^*)(\boldsymbol{\omega})|^2 + \lambda_n \frac{|\mathcal{F}(f_n^*)(\boldsymbol{\omega})|^2}{\mathcal{F}(\tilde{K}_S(\boldsymbol{x}))(\boldsymbol{\omega})} d\boldsymbol{\omega} \right) \\
\leq & C_1 \left(\int_{\mathbb{R}^D} |\mathcal{F}(f^*)(\boldsymbol{\omega}) - \mathcal{F}(\tilde{g}_n^*)(\boldsymbol{\omega})|^2 + \lambda_n \frac{|\mathcal{F}(\tilde{g}_n^*)(\boldsymbol{\omega})|^2}{\mathcal{F}(\tilde{K}_S(\boldsymbol{x}))(\boldsymbol{\omega})} d\boldsymbol{\omega} \right) \\
\leq & C_1 \left(\int_{\mathbb{R}^D} |\mathcal{F}(f^*)(\boldsymbol{\omega}) - \mathcal{F}(\tilde{g}_n^*)(\boldsymbol{\omega})|^2 + C_2\lambda_n |\mathcal{F}(\tilde{g}_n^*)(\boldsymbol{\omega})|^2 \sigma_n^{-2m_0} e^{3\sigma_n^2 \boldsymbol{\omega}^T \boldsymbol{\omega}} d\boldsymbol{\omega} \right) \\
= & C_1 \left(\int_{\mathbb{R}^D} \frac{C_2\lambda_n \sigma_n^{-2m_0} e^{3\sigma_n^2 \boldsymbol{\omega}^T \boldsymbol{\omega}}}{1 + C_2\lambda_n \sigma_n^{-2m_0} e^{3\sigma_n^2 \boldsymbol{\omega}^T \boldsymbol{\omega}}} |\mathcal{F}(f^*)(\boldsymbol{\omega})|^2 d\boldsymbol{\omega} \right) \\
\leq & C_3 \left(\int_{\Omega_1} \lambda_n \sigma_n^{-2m_0} e^{3\sigma_n^2 \boldsymbol{\omega}^T \boldsymbol{\omega}} |\mathcal{F}(f^*)(\boldsymbol{\omega})|^2 d\boldsymbol{\omega} + \int_{\Omega_1^C} |\mathcal{F}(f^*)(\boldsymbol{\omega})|^2 d\boldsymbol{\omega} \right) \\
= & C_3 \left(I_1 + I_2 \right),
\end{aligned}$$

for some positive constants C_1, C_2 and C_3, where $\tilde{g}_n^*$ minimizes

$$\int_{\mathbb{R}^D} |\mathcal{F}(f^*)(\boldsymbol{\omega}) - \mathcal{F}(g)(\boldsymbol{\omega})|^2 + C_2\lambda_n |\mathcal{F}(g)(\boldsymbol{\omega})|^2 \sigma_n^{-2m_0} e^{3\sigma_n^2 \boldsymbol{\omega}^T \boldsymbol{\omega}} d\boldsymbol{\omega},$$

$\Omega_1 = \{\boldsymbol{\omega} : C_2\lambda_n \sigma_n^{-2m_0} e^{3\sigma_n^2 \boldsymbol{\omega}^T \boldsymbol{\omega}} \leq 1\}$, which is the same as $\Omega_1 = \{\boldsymbol{\omega} : \|\boldsymbol{\omega}\|_2^2 < \frac{2m_0 \log \sigma_n - \log(C_2\lambda_n)}{3\sigma_n^2}\}$, provided that $C_2\lambda_n \sigma_n^{-2m_0} < 1$, and the third inequality is because of (151).

Let $g(u) = 3\sigma_n^2 u - m_f \log(1+u)$. Taking the derivative, we obtain

$$g'(u) = 3\sigma_n^2 - \frac{m_f}{1+u},$$

which is smaller than zero when $u \in [0, \frac{m_f}{3\sigma_n^2} - 1)$, and larger than zero when $u \in (\frac{m_f}{3\sigma_n^2} - 1, \infty)$.

Since $g(0) = 0$ and

$$\begin{aligned}
& g\left(\frac{2m_0 \log \sigma_n - \log(C_2\lambda_n)}{3\sigma_n^2} \right) = 2m_0 \log \sigma_n - \log(C_2\lambda_n) - m_f \log \left(1 + \frac{2m_0 \log \sigma_n - \log(C_2\lambda_n)}{3\sigma_n^2} \right) \\
\leq & 2m_0 \log \sigma_n - \log(C_2\lambda_n) - m_f \log \left((2m_0 \log \sigma_n - \log(C_2\lambda_n))/3 \right) + 2m_f \log \sigma_n \\
\leq & (2m_0 + 2m_f) \log \sigma_n - \log(C_2\lambda_n),
\end{aligned}$$

where the last inequality is because $C_2\lambda_n \sigma_n^{-2m_0} = o(1)$, which implies $\log\left((2m_0 \log \sigma_n - \log(C_2\lambda_n))/3 \right) > 0$ as n becomes large.

Therefore, for $u \in [0, \frac{2m_0 \log \sigma_n - \log(C_2\lambda_n)}{3\sigma_n^2}]$, we have

$$g(u) \leq \max(0, \log(\sigma_n^{(2m_0+2m_f)} (C_2\lambda_n)^{-1})),$$

which implies

$$e^{3\sigma_n^2\|\boldsymbol{\omega}\|_2^2} \leq \max(1, \sigma_n^{(2m_0+2m_f)}(C_2\lambda_n)^{-1})(1+\|\boldsymbol{\omega}\|_2^2)^{m_f},$$

for $\boldsymbol{\omega} \in \Omega_1$. Thus, the term I_1 can be bounded by

$$I_1 \leq \max(\lambda_n\sigma_n^{-2m_0}, C_2^{-1}\sigma_n^{2m_f}) \int_{\Omega_1} (1+|\boldsymbol{\omega}|^2)^{m_f}|\mathcal{F}(f^*)(\boldsymbol{\omega})|^2 d\boldsymbol{\omega}. \tag{154}$$

The term I_2 can be bounded by

$$\begin{aligned} I_2 \leq & \frac{3\sigma_n^{2m_f}}{(2m_0\log\sigma_n - \log(C_2\lambda_n))^{m_f}} \int_{\Omega_1^C} (1+|\boldsymbol{\omega}|^2)^{m_f}|\mathcal{F}(f^*)(\boldsymbol{\omega})|^2 d\boldsymbol{\omega} \\ \leq & 3C_4\sigma_n^{2m_f} \int_{\Omega_1^C} (1+|\boldsymbol{\omega}|^2)^{m_f}|\mathcal{F}(f^*)(\boldsymbol{\omega})|^2 d\boldsymbol{\omega}, \end{aligned} \tag{155}$$

for some positive constants C_4, where the first inequality is because on Ω_1^C, we have $\|\boldsymbol{\omega}\|^2 \geq \frac{2m_0\log\sigma_n - \log(C_2\lambda_n)}{3\sigma_n^2}$, which implies for sufficiently large n,

$$(1+\|\boldsymbol{\omega}\|_2^2)^{m_f} \geq \frac{(2m_0\log\sigma_n - \log(C_2\lambda_n))^{m_f}}{3\sigma_n^{2m_f}},$$

and the last inequality is because $C_2\lambda_n\sigma_n^{-2m_0} = o(1)$. Combining (154) and (155) leads to

$$\begin{aligned} I_1 + I_2 \leq & C_5\max(\lambda_n\sigma_n^{-2m_0}, \sigma_n^{2m_f}) \int_{\mathbb{R}^D} (1+|\boldsymbol{\omega}|^2)^{m_f}|\mathcal{F}(f^*)(\boldsymbol{\omega})|^2 d\boldsymbol{\omega} \\ \leq & C_6\max(\lambda_n\sigma_n^{-2m_0}, \sigma_n^{2m_f})\|f^*\|^2_{\mathcal{W}^{m_f}(\Omega)}, \end{aligned}$$

for some positive constants C_5 and C_6, which finishes the proof. ∎

J.3 Proof of Lemma 27

We first present a lemma used in this proof, which states the entropy numbers of RKHSs generated by the Gaussian kernels. Lemma 35 is an intermediate step of the proof of Theorem A.2 of Hamm and Steinwart (2021a). Lemma 36 is a direct result of the proof of Lemma 8.4 of van de Geer (2000) and Lemma 35.

Lemma 35 *Let $4\sigma^2 \leq 1$. Then for all $0 < p < 2$, there exists a constant $C_1 > 0$ only depending on D such that for all $\delta > 0$, we have*

$$H(\delta, \mathcal{B}_{\mathcal{H}_\sigma(\Omega)}, \|\cdot\|_{L_\infty(\Omega)}) \leq C_1\sigma^{-d}p^{-D-1}\delta^{-p}.$$

Lemma 36 *Suppose conditions of Theorem 9 are fulfilled. Then for some constant $C_2 > 0$ only related to the Assumption 1 and for $\delta > 0$ with*

$$\sqrt{n}\delta > 2C_2\max\left(\int_0^1 H(u, \mathcal{B}_{\mathcal{H}_\sigma(\Omega)}, \|\cdot\|_{L_\infty(\Omega)})^{1/2}\mathrm{d}u, 1\right),$$

we have for all $0 < p < 2$

$$\mathbb{P}\left(\sup_{g\in\mathcal{B}_{\mathcal{H}_\sigma(\Omega)}} \frac{\langle g, \boldsymbol{\epsilon}\rangle_n}{\|g\|_n^{1-\frac{p}{2}}} \geq \delta\right) \leq C_2 p^{-1} \exp\left(-\frac{n\delta^2}{C_2}\right).$$

Proof of Lemma 36. In order to characterize the role of p in Lemma 36, we note that in the last step of the proof of Lemma 8.4 of van de Geer (2000), we use

$$\begin{aligned}
&\sum_{s=1}^{\infty} C_2 \exp\left(-\frac{n\delta^2}{16C_2^2}2^{sp}\right) \leq \sum_{s=1}^{\infty} C_2 \exp\left(-\frac{n\delta^2}{16C_2^2}e^{sp/2}\right)\\
\leq &\sum_{s=1}^{\infty} C_2 \exp\left(-\frac{n\delta^2}{16C_2^2}(1+\frac{sp}{2})\right) = C_2 \exp\left(-\frac{n\delta^2}{16C_2^2}\right)\frac{\exp\left(-\frac{np\delta^2}{32C_2^2}\right)}{1-\exp\left(-\frac{np\delta^2}{32C_2^2}\right)}\\
\leq &\frac{32C_2^3}{np\delta^2}\exp\left(-\frac{n\delta^2}{16C_2^2}\right) \leq \frac{8C_2}{p}\exp\left(-\frac{n\delta^2}{16C_2^2}\right),
\end{aligned}$$

where the second and the third inequalities are by $e^u > 1+u$ for all $u \in \mathbb{R}$, and the last inequality is by $n\delta^2 > 4C_2^2$.

Then if $C_2 \geq 1$,

$$\frac{8C_2}{p}\exp\left(-\frac{n\delta^2}{16C_2^2}\right) \leq \frac{16C_2^2}{p}\exp\left(-\frac{n\delta^2}{16C_2^2}\right).$$

and if $0 < C_2 < 1$,

$$\frac{8C_2}{p}\exp\left(-\frac{n\delta^2}{16C_2^2}\right) \leq \frac{16C_2}{p}\exp\left(-\frac{n\delta^2}{16C_2}\right).$$

The rest of the proof is similar to the proof of Lemma 8.4 of van de Geer (2000). ■

Proof of Lemma 27. Since $\hat{f}$ is the solution to the optimization problem (69), we have that

$$\|\hat{f} - \boldsymbol{y}\|_n^2 + \lambda_n\|\hat{f}\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)}^2 \leq \|f_n^* - \boldsymbol{y}\|_n^2 + \lambda_n\|f_n^*\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)}^2, \tag{156}$$

where f_n^* is as in Lemma 26. By rearrangement, (156) implies

$$\|f - \hat{f}_n\|_n^2 + \lambda_n\|\hat{f}\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)}^2 \leq \|f - f_n^*\|_n^2 + \lambda_n\|f_n^*\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)}^2 + 2\langle\boldsymbol{\epsilon}, \hat{f} - f_n^*\rangle_n.$$

Theorem 10.46 of Wendland (2004) states that every RKHS defined on Ω possesses a natural extension to $\mathbb{R}^D$ with equivalent norms. Applying this natural extension to $\mathcal{H}_{\tilde{K}_S}(\Omega)$, we obtain that

$$\|f - \hat{f}_n\|_n^2 + C_3\lambda_n\|\hat{f}\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)}^2 \leq \|f - f_n^*\|_n^2 + C_4\lambda_n\|f_n^*\|_{\mathcal{H}_{\tilde{K}_S}(\mathbb{R}^D)}^2 + 2\langle\boldsymbol{\epsilon}, \hat{f} - f_n^*\rangle_n \tag{157}$$

for some positive constants C_3 and C_4. By assumption, we have

$$\|f - f_n^*\|_{L_2(P_{\mathbf{X}})}^2 + \lambda_n\|f_n^*\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)}^2 \leq T.$$

Then Lemma 26 implies $\|f_n^*\|^2_{\mathcal{H}_{\tilde{K}_S}(\mathbb{R}^D)} = O(\lambda_n^{-1}T)$. Taking $p = (\log n)^{-1} \in (0,2)$ and $\delta_n = C_5\sigma_n^{-d/2}p^{-(D+1)/2}n^{-1/2}$ (where C_5 is a constant only depending on D), we have $\sqrt{n}\delta_n = C_5\sigma_n^{-d/2}p^{-(D+1)/2}$. Applying Lemma 36, we obtain that with probability at least

$$C_6(\log n)\exp(-C_6^{-1}C_5^2\sigma_n^{-2d}p^{-2D-2}),$$

for some positive constants C_6, we have

$$2\langle \boldsymbol{\epsilon}, \hat{f} - f_n^*\rangle_n \leq C_7\|\hat{f} - f_n^*\|_n^{1-\frac{p}{2}}(\|\hat{f}\|_{\mathcal{H}_{\sigma_n/\sqrt{2}}(\Omega)} + \|f_n^*\|_{\mathcal{H}_{\sigma_n/\sqrt{2}}(\Omega)})^{\frac{p}{2}}C_5\sigma_n^{-d/2}p^{-(D+1)/2}n^{-1/2}, \tag{158}$$

for some positive constants C_7. Plugging (158) into (157) yields

$$\begin{aligned}
&\|f - \hat{f}\|_n^2 + \lambda_n\|\hat{f}\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)}\\
\leq&\|f - f_n^*\|_n^2 + \lambda_n\|f_n^*\|^2_{\mathcal{H}_{\tilde{K}_S}(\mathbb{R}^D)}\\
&+ C_8\sigma_n^{-d/2}p^{-(D+1)/2}n^{-1/2}\|\hat{f} - f_n^*\|_n^{1-\frac{p}{2}}(\|\hat{f}\|_{\mathcal{H}_{\sigma_n/\sqrt{2}}(\Omega)} + \|f_n^*\|_{\mathcal{H}_{\sigma_n/\sqrt{2}}(\Omega)})^{\frac{p}{2}},
\end{aligned} \tag{159}$$

for some positive constants C_8. Now we consider bounding the difference between $\|f - f_n^*\|_n$ and $\|f - f_n^*\|_{L_2(P_{\mathbf{X}})}$. Since f_n^* does not depend on $\boldsymbol{x}_j$ and $\boldsymbol{\epsilon}$, we can directly apply Lemma 33 to $\|f - f_n^*\|_n$ and obtain that

$$\left|\|f - f_n^*\|_n^2 - \|f - f_n^*\|^2_{L_2(P_{\mathbf{X}})}\right| = O_{\mathbb{P}}(n^{-1/2})\|f - f_n^*\|_{L_2(P_{\mathbf{X}})},$$

which, together with Lemma 26, yields

$$\|f - f_n^*\|_n^2 = O_{\mathbb{P}}\left(T + n^{-1/2}T^{1/2}\right). \tag{160}$$

Plugging (160) into (159), together with Lemma 26, gives us

$$\begin{aligned}
&\|f - \hat{f}\|_n^2 + \lambda_n\|\hat{f}\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)}\\
\leq&O_{\mathbb{P}}\left(T + n^{-1/2}T^{1/2}\right)\\
&+ C_8\sigma_n^{-d/2-\frac{pD}{4}}p^{-(D+1)/2}n^{-1/2}\|\hat{f} - f_n^*\|_n^{1-\frac{p}{2}}(\|\hat{f}\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)} + \|f_n^*\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)})^{\frac{p}{2}},
\end{aligned} \tag{161}$$

where we also use $\sigma_n^{-D/2}\|f_n^*\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)} \geq C_8\|f_n^*\|_{\mathcal{H}_{\sigma_n/\sqrt{2}}(\Omega)}$. Then (161) implies either

$$\|f - \hat{f}\|_n^2 + \lambda_n\|\hat{f}\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)} = O_{\mathbb{P}}\left(T + n^{-1/2}T^{1/2}\right), \tag{162}$$

or

$$\begin{aligned}
&\|f - \hat{f}\|_n^2 + \lambda_n\|\hat{f}\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)}\\
\leq&4C_8\sigma_n^{-d/2-\frac{pD}{4}}p^{-(D+1)/2}n^{-1/2}\|\hat{f} - f_n^*\|_n^{1-\frac{p}{2}}(\|\hat{f}\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)} + \|f_n^*\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)})^{\frac{p}{2}},
\end{aligned} \tag{163}$$

In order to solve (163), we consider two cases.

Case 1: $\|\hat{f}\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)} \geq \|f_n^*\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)}$. In this case, we have

$$
\begin{aligned}
&\|f-\hat{f}\|_n^2 + \lambda_n \|\hat{f}\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)} \leq 8C_8 \sigma_n^{-d/2-\frac{pD}{4}} p^{-(D+1)/2} n^{-1/2} \|\hat{f} - f_n^*\|_n^{1-\frac{p}{2}} \|\hat{f}\|^{\frac{p}{2}}_{\mathcal{H}_{\tilde{K}_S}(\Omega)} \\
\leq& 8C_8 \sigma_n^{-d/2-\frac{pD}{4}} p^{-(D+1)/2} n^{-1/2} \|f - f_n^*\|_n^{1-\frac{p}{2}} \|\hat{f}\|^{\frac{p}{2}}_{\mathcal{H}_{\tilde{K}_S}(\Omega)} \\
&+ 8C_8 \sigma_n^{-d/2-\frac{pD}{4}} p^{-(D+1)/2} n^{-1/2} \|f - \hat{f}\|_n^{1-\frac{p}{2}} \|\hat{f}\|^{\frac{p}{2}}_{\mathcal{H}_{\tilde{K}_S}(\Omega)},
\end{aligned} \tag{164}
$$

where the second equality is because of the basic inequality $(a+b)^q \leq a^q + b^q$ for $q \in (0,1)$. It can be seen that (164) further implies

$$
\|f-\hat{f}\|_n^2 + \lambda_n \|\hat{f}\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)} \leq 8C_8 \sigma_n^{-d/2-\frac{pD}{4}} p^{-(D+1)/2} n^{-1/2} \|f - f_n^*\|_n^{1-\frac{p}{2}} \|\hat{f}\|^{\frac{p}{2}}_{\mathcal{H}_{\tilde{K}_S}(\Omega)}, \tag{165}
$$

or

$$
\|f-\hat{f}\|_n^2 + \lambda_n \|\hat{f}\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)} \leq 8C_8 \sigma_n^{-d/2-\frac{pD}{4}} p^{-(D+1)/2} n^{-1/2} \|f - \hat{f}\|_n^{1-\frac{p}{2}} \|\hat{f}\|^{\frac{p}{2}}_{\mathcal{H}_{\tilde{K}_S}(\Omega)}. \tag{166}
$$

Solving (166) yields

$$
\begin{aligned}
\|f-\hat{f}\|_n \leq& 8C_8 \sigma_n^{-d/2-\frac{pD}{4}} p^{-(D+1)/2} n^{-1/2} \lambda_n^{-\frac{p}{4}}, \\
\|\hat{f}\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)} \leq& 8C_8 \sigma_n^{-d/2-\frac{pD}{4}} p^{-(D+1)/2} n^{-1/2} \lambda_n^{-\frac{2+p}{4}}.
\end{aligned} \tag{167}
$$

Plugging (160) into (165), we have

$$
\|f-\hat{f}\|_n^2 + \lambda_n \|\hat{f}\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)} \leq 8C_8 \sigma_n^{-d/2-\frac{pD}{4}} p^{-(D+1)/2} n^{-1/2} (T + n^{-1/2} T^{1/2})^{\frac{1}{2}-\frac{p}{4}} \|\hat{f}\|^{\frac{p}{2}}_{\mathcal{H}_{\tilde{K}_S}(\Omega)}. \tag{168}
$$

Solving (168) yields

$$
\begin{aligned}
\|f-\hat{f}\|_n \leq& \lambda_n^{-\frac{p}{2(4-p)}} \left(8C_8 \sigma_n^{-d/2-\frac{pD}{4}} p^{-(D+1)/2} n^{-1/2} (T + n^{-1/2} T^{1/2})^{\frac{1}{2}-\frac{p}{4}} \right)^{\frac{2}{4-p}}, \\
\|\hat{f}\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)} \leq& \left(8C_8 \lambda_n^{-1} \sigma_n^{-d/2-\frac{pD}{4}} p^{-(D+1)/2} n^{-1/2} (T + n^{-1/2} T^{1/2})^{\frac{1}{2}-\frac{p}{4}} \right)^{\frac{2}{4-p}}.
\end{aligned} \tag{169}
$$

Case 2: $\|\hat{f}\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)} < \|f_n^*\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)}$. In this case, (163) implies that

$$
\begin{aligned}
&\|f-\hat{f}\|_n^2 + \lambda_n \|\hat{f}\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)} \leq 8C_8 \sigma_n^{-d/2-\frac{pD}{4}} p^{-(D+1)/2} n^{-1/2} \|\hat{f} - f_n^*\|_n^{1-\frac{p}{2}} \|f_n^*\|^{\frac{p}{2}}_{\mathcal{H}_{\tilde{K}_S}(\Omega)} \\
\leq& 8C_8 \sigma_n^{-d/2-\frac{pD}{4}} p^{-(D+1)/2} n^{-1/2} \|f - f_n^*\|_n^{1-\frac{p}{2}} \|f_n^*\|^{\frac{p}{2}}_{\mathcal{H}_{\tilde{K}_S}(\Omega)} \\
&+ 8C_8 \sigma_n^{-d/2-\frac{pD}{4}} p^{-(D+1)/2} n^{-1/2} \|f - \hat{f}\|_n^{1-\frac{p}{2}} \|f_n^*\|^{\frac{p}{2}}_{\mathcal{H}_{\tilde{K}_S}(\Omega)},
\end{aligned} \tag{170}
$$

where the second equality is because of the basic inequality $(a+b)^q \leq a^q + b^q$ for $q \in (0,1)$. By (170), we have either

$$\|f-\hat{f}\|_n^2 + \lambda_n \|\hat{f}\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)} \leq C_9 \sigma_n^{-d/2-\frac{pD}{4}} p^{-(D+1)/2} n^{-1/2} \|f - f_n^*\|_n^{1-\frac{p}{2}} \|f_n^*\|^{\frac{p}{2}}_{\mathcal{H}_{\tilde{K}_S}(\Omega)}, \tag{171}$$

or

$$\|f-\hat{f}\|_n^2 + \lambda_n \|\hat{f}\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)} \leq C_{10} \sigma_n^{-d/2-\frac{pD}{4}} p^{-(D+1)/2} n^{-1/2} \|f - \hat{f}\|_n^{1-\frac{p}{2}} \|f_n^*\|^{\frac{p}{2}}_{\mathcal{H}_{\tilde{K}_S}(\Omega)}, \tag{172}$$

for some positive constants C_9 and C_{10}. Combining (171) and Lemma 26, we have

$$\begin{aligned}
\|f-\hat{f}\|_n^2 =& O_{\mathbb{P}}\left(\sigma_n^{-d/2-\frac{pD}{4}} p^{-(D+1)/2} n^{-1/2} (\lambda_n^{-1}T)^{\frac{p}{2}} (T + n^{-1/2}T^{1/2})^{1-\frac{p}{2}}\right), \\
\|\hat{f}\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)} =& O_{\mathbb{P}}\left(\lambda_n^{-1}\sigma_n^{-d/2-\frac{pD}{4}} p^{-(D+1)/2} n^{-1/2} (\lambda_n^{-1}T)^{\frac{p}{2}} (T + n^{-1/2}T^{1/2})^{1-\frac{p}{2}}\right).
\end{aligned} \tag{173}$$

Combining (172) and Lemma 26, we have

$$\begin{aligned}
\|f-\hat{f}\|_n =& O_{\mathbb{P}}\left((\sigma_n^{-d/2-\frac{pD}{4}} p^{-(D+1)/2} n^{-1/2})^{\frac{2}{2+p}} (\lambda_n^{-1}T)^{\frac{p}{2+p}}\right) \\
\|\hat{f}\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)} =& O_{\mathbb{P}}\left(\lambda_n^{-1/2}(\sigma_n^{-d/2-\frac{pD}{4}} p^{-(D+1)/2} n^{-1/2})^{\frac{2}{2+p}} (\lambda_n^{-1}T)^{\frac{p}{2+p}}\right).
\end{aligned} \tag{174}$$

By (162), (167), (169), (173), and (174), we finish the proof. ■

J.4 Proof of Lemma 28

For any function $g \in \mathcal{H}_\sigma(\mathbb{R}^D)$, the Fourier inversion theorem implies

$$\begin{aligned}
|g(\boldsymbol{x})| =& \left|\int_{\mathbb{R}^D} e^{i\boldsymbol{x}^T\boldsymbol{\omega}} \mathcal{F}(g)(\boldsymbol{\omega}) d\boldsymbol{\omega}\right| \leq \int_{\mathbb{R}^D} |\mathcal{F}(g)(\boldsymbol{\omega})| \, d\boldsymbol{\omega} \\
=& \int_{\mathbb{R}^D} |\mathcal{F}(g)(\boldsymbol{\omega})|^{1-r} (\mathcal{F}(k_\sigma)(\boldsymbol{\omega}))^{r/2} |\mathcal{F}(g)(\boldsymbol{\omega})|^{r} (\mathcal{F}(k_\sigma)(\boldsymbol{\omega}))^{-r/2} d\boldsymbol{\omega} \\
\leq& \left(\int_{\mathbb{R}^D} |\mathcal{F}(g)(\boldsymbol{\omega})|^{\frac{2(1-r)}{2-r}} (\mathcal{F}(k_\sigma)(\boldsymbol{\omega}))^{\frac{r}{2-r}} d\boldsymbol{\omega}\right)^{\frac{2-r}{2}} \left(\int_{\mathbb{R}^D} |\mathcal{F}(g)(\boldsymbol{\omega})|^2 (\mathcal{F}(k_\sigma)(\boldsymbol{\omega}))^{-1} d\boldsymbol{\omega}\right)^{\frac{r}{2}} \\
\leq& 2^{\frac{Dr}{2}} \sigma^{\frac{Dr}{2}} \left(\int_{\mathbb{R}^D} |\mathcal{F}(g)(\boldsymbol{\omega})|^{\frac{2(1-r)}{2-r}} e^{-\frac{r}{2-r}\sigma^2\|\boldsymbol{\omega}\|_2^2} d\boldsymbol{\omega}\right)^{\frac{2-r}{2}} \|g\|^r_{\mathcal{H}_\sigma(\mathbb{R}^D)} \\
\leq& 2^{\frac{Dr}{2}} \sigma^{\frac{Dr}{2}} \left(\int_{\mathbb{R}^D} |\mathcal{F}(g)(\boldsymbol{\omega})|^2 d\boldsymbol{\omega}\right)^{\frac{1-r}{2}} \left(\int_{\mathbb{R}^D} e^{-r\sigma^2\|\boldsymbol{\omega}\|_2^2} d\boldsymbol{\omega}\right)^{\frac{1}{2}} \|g\|^r_{\mathcal{H}_\sigma(\mathbb{R}^D)} \\
=& 2^{\frac{Dr}{2}} \sigma^{\frac{Dr}{2}} (4\pi^{-1} r\sigma^2)^{-\frac{D}{4}} \|g\|^{1-r}_{L_2(\mathbb{R}^D)} \|g\|^r_{\mathcal{H}_\sigma(\mathbb{R}^D)} \\
\leq& C_1 r^{-\frac{D}{4}} \sigma^{\frac{D(r-1)}{2}} \|g\|^{1-r}_{L_2(\mathbb{R}^D)} \|g\|^r_{\mathcal{H}_\sigma(\mathbb{R}^D)},
\end{aligned}$$

for some positive constants C_1, where the second and fourth inequalities are by Hölder's inequality, and the third equality is by Parseval's identity. This finishes the proof. ■

Appendix K. Proof of Lemmas in Appendix F

K.1 Proof of Lemma 29

By following the similar approach in Appendix I.1, we have

$$\|f^* - f_n^*\|_{L_2(P_{\mathbf{X}})}^2 + \lambda_n \|f_n^*\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)}^2 \leq \max(C_1, 1) \left(\|f^* - \tilde{f}_n^*\|_{L_2(\mathbb{R}^D)}^2 + \lambda_n \|\tilde{f}_n^*\|_{\mathcal{H}_{\tilde{K}_S}(\mathbb{R}^D)}^2 \right). \tag{175}$$

Therefore, it remains to bound

$$\|f^* - \tilde{f}_n^*\|_{L_2(\mathbb{R}^D)}^2 + \lambda_n \|\tilde{f}_n^*\|_{\mathcal{H}_{\tilde{K}_S}(\mathbb{R}^D)}^2.$$

Similar to Appendix I.1, we can use the Fourier inversion theorem to get

$$\begin{aligned}
&\|f^* - \tilde{f}_n^*\|_{L_2(\mathbb{R}^D)}^2 + \lambda_n \|\tilde{f}_n^*\|_{\mathcal{H}_{\tilde{K}_S}(\mathbb{R}^D)}^2 = \int_{\mathbb{R}^D} |\mathcal{F}(f^*)(\boldsymbol{\omega}) - \mathcal{F}(\tilde{f}_n^*)(\boldsymbol{\omega})|^2 + \lambda_n \frac{|\mathcal{F}(\tilde{f}_n^*)(\boldsymbol{\omega})|^2}{\mathcal{F}(\tilde{K}_S(\boldsymbol{x}))(\boldsymbol{\omega})} \mathrm{d}\boldsymbol{\omega} \\
&\leq \int_{\mathbb{R}^D} \frac{C_2 \lambda_n \prod_{j=1}^D (1+\omega_j^2)^{m_0} (1+\sigma_n^2 w_j^2)^{m_\varepsilon}}{1 + C_2 \lambda_n \prod_{j=1}^D (1+\omega_j^2)^{m_0} (1+\sigma_n^2 w_j^2)^{m_\varepsilon}} |\mathcal{F}(f^*)(\boldsymbol{\omega})|^2 \mathrm{d}\boldsymbol{\omega} \\
&\leq \sum_{|\boldsymbol{l}| \geq 1} I_{\boldsymbol{l}}^{<} + I_{\boldsymbol{l}}^{\geq},
\end{aligned} \tag{176}$$

for some positive constants C_2, where $\boldsymbol{l} = (l_1, ..., l_D) \in \{0,1\}^D$,

$$\begin{aligned}
\Omega_{l_j} &= \begin{cases} \{\omega_j : \sigma_n^2 \omega_j^2 < 1\}, & \text{if } l_j = 0, \\ \{\omega_j : \sigma_n^2 \omega_j^2 \geq 1\}, & \text{otherwise}, \end{cases} \\
\Omega_{\boldsymbol{l}}^{<} &= \left[\times_{j=1}^D \Omega_{l_j}\right] \bigcap \{\boldsymbol{\omega} : C_2 \lambda_n \prod_{j=1}^D (1+\omega_j^2)^{m_0} (1+\sigma_n^2 w_j^2)^{m_\varepsilon} < 1\}, \\
\Omega_{\boldsymbol{l}}^{\geq} &= \left[\times_{j=1}^D \Omega_{l_j}\right] \bigcap \{\boldsymbol{\omega} : C_2 \lambda_n \prod_{j=1}^D (1+\omega_j^2)^{m_0} (1+\sigma_n^2 w_j^2)^{m_\varepsilon} \geq 1\}, \\
I_{\boldsymbol{l}}^{<} &= \int_{\Omega_{\boldsymbol{l}}^{<}} C_2 \lambda_n \left[\prod_{j=1}^D (1+\omega_j^2)^{m_0} (1+\sigma_n^2 w_j^2)^{m_\varepsilon} \right] |\mathcal{F}(f^*)(\boldsymbol{\omega})|^2 \mathrm{d}\boldsymbol{\omega}, \\
I_{\boldsymbol{l}}^{\geq} &= \int_{\Omega_{\boldsymbol{l}}^{\geq}} |\mathcal{F}(f^*)(\boldsymbol{\omega})|^2 \mathrm{d}\boldsymbol{\omega},
\end{aligned}$$

and the sum over all $\{|\boldsymbol{l}| \geq 1\}$ is because on any $\Omega_{\boldsymbol{l}}^{<}$ and $\Omega_{\boldsymbol{l}}^{\geq}$, there must be at least one j^* and one j^{**} such that $\sigma_n^2 w_{j^*} < 1$ and $\sigma_n^2 w_{j^{**}} \geq 1$, respectively.

Define $p = \frac{m_f}{m_0+m_\varepsilon} \le 1$. On any $\Omega_{\boldsymbol{l}}^{<}$, we have

$$
\begin{aligned}
&C_2\lambda_n \prod_{j=1}^{D}(1+\omega_j^2)^{m_0}(1+\sigma_n^2 w_j^2)^{m_\varepsilon}\\
\le&\left(C_2\prod_{j=1}^{D}\lambda_n^{\frac{1}{D}}(1+\omega_j^2)^{m_0}(1+\sigma_n^2 w_j^2)^{m_\varepsilon}\right)^p\\
=&C_3\prod_{j=1}^{D}\lambda_n^{\frac{p}{D}}(1+\omega_j^2)^{m_0 p}(1+\sigma_n^2 w_j^2)^{m_\varepsilon p}\\
\le&C_4\prod_{j=1}^{D}\left(\lambda_n^{\frac{p}{D}}(1+\omega_j^2)^{m_0 p}\right)^{1-l_j}\left(\lambda_n^{\frac{p}{D}}(1+\omega_j^2)^{m_0 p}(\sigma_n^2 w_j^2)^{m_\varepsilon p}\right)^{l_j},
\end{aligned}
$$

for some positive constants C_3 and C_4.
From the fact that $m_0 p = m_f\frac{m_0}{m_0+m_\varepsilon} \le m_f$ and calculations similar to (125), we have

$$
\begin{aligned}
&\lambda_n^{\frac{p}{D}}(1+\omega_j^2)^{m_0 p} \le \lambda_n^{\frac{p}{D}}(1+\omega_j^2)^{m_f} && \text{when } l_j = 0,\\
\text{and}\quad &\lambda_n^{\frac{p}{D}}(1+\omega_j^2)^{m_0 p}(\sigma_n^2 w_j^2)^{m_\varepsilon p} \le (\lambda_n^{\frac{1}{D}}\sigma_n^{2m_\varepsilon})^p(1+\omega_j^2)^{m_f} && \text{when } l_j = 1.
\end{aligned}
$$

As a result, on $\Omega_{\boldsymbol{l}}^{<}$, we have

$$
\begin{aligned}
C_2\lambda_n \prod_{j=1}^{D}(1+\omega_j^2)^{m_0}(1+\sigma_n^2 w_j^2)^{m_\varepsilon} &\le C_4\prod_{j=1}^{D}\left(\lambda_n^{\frac{p}{D}}\right)^{1-l_j}\left(\lambda_n^{\frac{1}{D}}\sigma_n^{2m_\varepsilon}\right)^{pl_j}(1+\omega_j^2)^{m_f}\\
&= C_4\lambda_n^p\sigma_n^{2m_\varepsilon p|\boldsymbol{l}|}\prod_{j=1}^{D}(1+\omega_j^2)^{m_f}, \qquad (177)
\end{aligned}
$$

where $|\boldsymbol{l}| = \sum_{j=1}^{D} l_j$.

On $\Omega_{\boldsymbol{l}}^{\ge}$, we have

$$
\begin{aligned}
1 \le& \left(C_2\lambda_n\prod_{j=1}^{D}(1+\omega_j^2)^{m_0}(1+\sigma_n^2\omega_j^2)^{m_\varepsilon}\right)^p\\
\le& \left(C_5\lambda_n\prod_{j=1}^{D}(1+\omega_j^2)^{m_0}(\sigma_n^2\omega_j^2)^{m_\varepsilon l_j}\right)^p\\
\le& C_6\lambda_n^p\sigma_n^{2m_\varepsilon p|\boldsymbol{l}|}\prod_{j=1}^{D}(1+\omega_j)^{m_f}, \qquad (178)
\end{aligned}
$$

for some positive constants C_5 and C_6.
Plugging (177) and (178) into (176) finishes the proof. ■

K.2 Proof of Lemma 30

Let $\Psi_\sigma(\|\cdot\|_2) \coloneqq \prod_{j=1}^D \psi_\sigma(|\cdot|)$ be tensor product of positive definite functions with

$$c_1(1+\sigma^2|\omega_j|^2)^{-(m_0+m_\varepsilon)} \leq \mathcal{F}(\psi_\sigma) \leq c_2(1+\sigma^2|\omega_j|^2)^{-(m_0+m_\varepsilon)}, \forall \boldsymbol{\omega} \in \mathbb{R}^D, \forall j = 1, \ldots, d$$

for some positive constants c_1 and c_2, and $\mathcal{N}_\sigma(\Omega)$ be the RKHS generated by Ψ_σ. We will use the following lemmas. Lemma 37 can be derived by Corollary A.8 of Hamm and Steinwart (2021a) and (6.6) of Dung et al. (2018). Lemma 38 is a direct result of the proof of Lemma 8.4 of van de Geer (2000) and Lemma 37.

Lemma 37 *Let $4\sigma^2 \leq 1$. Suppose the conditions of Lemma 30 are fulfilled. Let $\mathcal{B}_{\mathcal{N}_\sigma(\Omega)}$ be a unit ball in $\mathcal{N}_\sigma(\Omega)$. Then there exists a constant $C_1 > 0$ only depending on D and Ω such that for all $\delta > 0$, we have*

$$H(\delta, \mathcal{B}_{\mathcal{H}_\sigma(\Omega)}, \|\cdot\|_{L_\infty(\Omega)}) \leq C_1 \sigma^{-d} \delta^{-\frac{1}{m_0+m_\varepsilon}} |\log\delta|^{(D-1)+\frac{1}{2(m_0+m_\varepsilon)}}.$$

Lemma 38 *Suppose conditions of Theorem 13 are fulfilled. Then for any T large enough we have*

$$\mathbb{P}\left(\sup_{g\in\mathcal{B}_{\mathcal{H}_\sigma(\Omega)}} \frac{\sqrt{n}\langle g, \boldsymbol{\epsilon}\rangle_n}{\|g\|_n^{1-p}|\log\|g\|_n|^{(D-1+p)/2}} \geq T\right) \leq C_2 \exp\left(-\frac{T^2}{C_3}\right).$$

where $p = \frac{1}{2(m_0+m_\varepsilon)}$, C_2 and C_3 are some constant independent of T and n.

Proof of Lemma 38. From Lemma 37, we can derive that for any $\delta \leq 1$,

$$\int_0^\delta H(u, \mathcal{B}_{\mathcal{H}_\sigma(\Omega)}, \|\cdot\|_{L_\infty(\Omega)})^{1/2} \mathrm{d}u \lesssim \sigma^{-d}\delta^{1-p}\left|\log\delta\right|^{\frac{D-1+p}{2}}.$$

Then, by Corollary 8.3 of van de Geer (2000), we can derive that

$$\mathbb{P}\left(\sup_{g\in\mathcal{B}_{\mathcal{H}_\sigma(\Omega)}} \sqrt{n}\left|\langle g, \boldsymbol{\epsilon}\rangle_n\right| \geq \sigma^{-d}\delta^{1-p}\left|\log\delta\right|^{\frac{D-1+p}{2}}\right) \lesssim \exp\left(-C_4\sigma_n^{-2d}\delta^{-2p}|\log\delta|^{D-1+p}\right).$$

We then can follow the peeling-off argument in Lemma 8.4 of van de Geer (2000) to show

$$\begin{aligned}
&\mathbb{P}\left(\sup_{g\in\mathcal{B}_{\mathcal{H}_\sigma(\Omega)}} \frac{\sqrt{n}\langle g, \boldsymbol{\epsilon}\rangle_n}{\|g\|_n^{1-p}|\log\|g\|_n|^{(D-1+p)/2}} \geq T\right)\\
\leq &\sum_{s=1}^\infty \mathbb{P}\left(\sup_{g\in\mathcal{B}_{\mathcal{H}_\sigma(\Omega)}, \|g\|_n\leq 2^{-s+1}} \sqrt{n}\langle g, \boldsymbol{\epsilon}\rangle_n \geq T2^{-s(1-p)}s^{\frac{D-1+p}{2}}\right)\\
\lesssim &\sum_{s=1}^\infty \exp\left(-C_4T^2\sigma_n^{-2d}2^{4ps}|\log 2|^{D-1+p}\right)\\
\lesssim &\sum_{s=1}^\infty \exp\left(-C_4T^2s\right)\\
= &C_2\exp\left(-\frac{T^2}{C_3}\right),
\end{aligned}$$

for some positive constants C_4. ■

Proof of Lemma 30. We can follow the proof of Lemma 22 to derive the following inequality using Lemmas 37 and 38:

$$\begin{aligned}
&\|f^* - \hat{f}_n\|_n^2 + \lambda_n \|\hat{f}_n\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)} \\
\leq &\|f^* - f_n^*\|_n^2 + \lambda_n \|f_n^*\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)} \\
&\quad + O_{\mathbb{P}}(n^{-1/2})\sigma_n^{-d/2}\|\hat{f}_n - f_n^*\|_n^{1-\frac{p}{2}} |\log \|\hat{f}_n - f_n^*\|_n|^{\frac{D-1}{2}+\frac{p}{4}} (\|\hat{f}_n\|_{\mathcal{N}_{\sigma_n}(\Omega)} + \|f_n^*\|_{\mathcal{N}_{\sigma_n}(\Omega)})^{\frac{p}{2}},
\end{aligned} \tag{179}$$

where $p = \frac{1}{m_0 + m_\varepsilon}$. Notice that (179) is similar to (133) in the proof of Lemma 22 except for the extra poly-log term $|\log \|\hat{f}_n - f_n^*\|_n|^{\frac{D-1}{2}+\frac{p}{4}}$. However, the extra poly-log term will not change the case-by-case analysis in our proof because it is always dominated by those polynomial terms in (179). Therefore, we can follow the same logic in the proof of Lemma 22 to get the final results.

K.3 Proof of Lemma 31

For any function $g \in \mathcal{MW}^m(\mathbb{R}^D)$, the Fourier inversion theorem implies

$$\begin{aligned}
|g(\boldsymbol{x})| = &\left| \int_{\mathbb{R}^D} e^{i\boldsymbol{x}^T\boldsymbol{\omega}} \mathcal{F}(g)(\boldsymbol{\omega}) d\boldsymbol{\omega} \right| \leq \int_{\mathbb{R}^D} |\mathcal{F}(g)(\boldsymbol{\omega})| \, d\boldsymbol{\omega} \\
= &\int_{\mathbb{R}^D} |\mathcal{F}(g)(\boldsymbol{\omega})|^{1-r} \left(\mathcal{F}(k_\sigma)(\boldsymbol{\omega})\right)^{r/2} |\mathcal{F}(g)(\boldsymbol{\omega})|^r \left(\mathcal{F}(k_\sigma)(\boldsymbol{\omega})\right)^{-r/2} d\boldsymbol{\omega} \\
\leq &\left(\int_{\mathbb{R}^D} |\mathcal{F}(g)(\boldsymbol{\omega})|^{\frac{2(1-r)}{2-r}} \left(\mathcal{F}(k_\sigma)(\boldsymbol{\omega})\right)^{\frac{r}{2-r}} d\boldsymbol{\omega} \right)^{\frac{2-r}{2}} \left(\int_{\mathbb{R}^D} |\mathcal{F}(g)(\boldsymbol{\omega})|^2 \left(\mathcal{F}(k_\sigma)(\boldsymbol{\omega})\right)^{-1} d\boldsymbol{\omega} \right)^{\frac{r}{2}} \\
\leq &\left(\int_{\mathbb{R}^D} |\mathcal{F}(g)(\boldsymbol{\omega})|^{\frac{2(1-r)}{2-r}} \left| \prod_{j=1}^D (1+\omega_j^2)^{-m} \right|^{\frac{r}{2-r}} d\boldsymbol{\omega} \right)^{\frac{2-r}{2}} \|g\|^r_{\mathcal{H}_\sigma(\mathbb{R}^D)} \\
\leq &\left(\int_{\mathbb{R}^D} |\mathcal{F}(g)(\boldsymbol{\omega})|^2 d\boldsymbol{\omega} \right)^{\frac{1-r}{2}} \left(\prod_{j=1}^D \int_{\mathbb{R}} (1+\omega_j^2)^{-mr} d\boldsymbol{\omega} \right)^{\frac{1}{2}} \|g\|^r_{\mathcal{H}_\sigma(\mathbb{R}^D)} \\
= &C_r \|g\|^{1-r}_{L_2(\mathbb{R}^D)} \|g\|^r_{\mathcal{H}_\sigma(\mathbb{R}^D)},
\end{aligned}$$

where the second and fourth inequalities are by Hölder's inequality, and the third equality is by Parseval's identity. The positive constant $C_r < \infty$ for any $r > m^{-1}/2$. This finishes the proof. ■

Appendix L. Proof of Lemma 32

Lemma 39 *Let the RKHS $\mathcal{H}_m$ induced by the kernel function K_m be equipped with norm satisfying*

$$\|f\|_{\mathcal{H}_m}^2 \le C \int_{\mathbb{R}^D} \left(1 + \frac{\|\boldsymbol{\omega}\|^2}{m}\right)^m \left|\hat{f}(\boldsymbol{\omega})\right|^2 \mathrm{d}\boldsymbol{\omega},$$

where C is some constant independent of m. Then for any $m > D/2$, there exists a constant C' independent of m such that for all $\delta > 0$, we have

$$H(\delta, \mathcal{B}_{\mathcal{H}_m([0,1]^D)}, \|\cdot\|_{L_\infty([0,1]^D)}) \le C'(2m-D)^{-\frac{2D}{2m-D}} m^{\frac{2mD}{2m-D}} \delta^{-\frac{2D}{2m-D}} \log(1+\delta^{-1}).$$

Remark 40 *If we treat m as a constant, the upper bound in Lemma 39 is larger than that in* (130)*. However, in the proofs of Lemmas 32 and 22, it turns out that the upper bound in Lemma 39 is sufficient.*

Proof of Lemma 32. The proof follows Corollary A.8 of Hamm and Steinwart (2021a). Specifically, Corollary A.8 of Hamm and Steinwart (2021a) states that for any $\delta > 0$ annd $\sigma > 0$, it holds that

$$H(\delta, \mathcal{B}_{\mathcal{H}_\sigma(\Omega)}, \|\cdot\|_{L_\infty(\Omega)}) \le \mathcal{N}_{\ell_\infty^D}(\sigma, \Omega) H(\delta, \mathcal{B}_{\mathcal{H}_m([0,1]^D)}, \|\cdot\|_{L_\infty(\Omega)}),$$

which, by Assumption 5 and Lemma 39, leads to

$$H(\delta, \mathcal{B}_{\mathcal{H}_\sigma(\Omega)}, \|\cdot\|_{L_\infty(\Omega)}) \le C\sigma^{-d}(2m-D)^{-\frac{2D}{2m-D}} m^{\frac{2mD}{2m-D}} \delta^{-\frac{2D}{2m-D}} \log(1+\delta^{-1}),$$

where the constant C is independent with m_ε, and $m = m_\varepsilon + m_0$. ■

Appendix M. Proof of Lemma 39

For any $f \in \mathcal{H}_m([0,1]^D)$, we have the following representation of f by Fourier series

$$f = \sum_{\boldsymbol{\zeta} \in \mathbb{N}^D} f_{\boldsymbol{\zeta}} \psi_{\boldsymbol{\zeta}},$$

where $\psi_{\boldsymbol{\zeta}}$ is the Fourier basis associated to $\boldsymbol{\zeta}$ and $f_{\boldsymbol{\zeta}}$ is the projection of f on $\psi_{\boldsymbol{\zeta}}$. Then transference from $L_2(\mathbb{R}^D)$ to $L_2([0,1]^D)$ by Fourier multiplier (see theorem 3.4 in L Coifman and Weiss (1977)) shows that the RKHS norm of f embedded on $[0,1]^D$ can be written as

$$\|f\|_{\mathcal{H}_m}^2 \le \sum_{\boldsymbol{\zeta} \in \mathbb{N}^D} (1 + \frac{\|\boldsymbol{\zeta}\|_2^2}{m})^m f_{\boldsymbol{\zeta}}^2.$$

We first define a projection P_M as follows:

$$P_M f = \sum_{\boldsymbol{\zeta} \in [M]^D} f_{\boldsymbol{\zeta}} \psi_{\boldsymbol{\zeta}}.$$

Then for the embedding operator $\mathcal{I}:\mathcal{H}([0,1]^D)\to L_\infty([0,1]^D)$, we have

$$\begin{aligned}\|\mathcal{I}\|_2^2 &= \sup_{f\in\mathcal{B}_{\mathcal{H}_m([0,1]^D)}}\sup_{\boldsymbol{x}\in[0,1]^D}|f(\boldsymbol{x})|^2\\ &\leq 2\sup_{f\in\mathcal{B}_{\mathcal{H}_m([0,1]^D)}}\sup_{\boldsymbol{x}\in[0,1]^D}|P_Mf(\boldsymbol{x})|^2+2\sup_{f\in\mathcal{B}_{\mathcal{H}_m([0,1]^D)}}\sup_{\boldsymbol{x}\in[0,1]^D}|f(\boldsymbol{x})-P_Mf(\boldsymbol{x})|^2.\end{aligned}\tag{180}$$

For the first term of (180), it is obvious that

$$2\sup_{f\in\mathcal{B}_{\mathcal{H}_m([0,1]^D)}}\sup_{\boldsymbol{x}\in[0,1]^D}|P_Mf(\boldsymbol{x})|^2\leq 2\|\mathcal{I}\|_2^2\leq 2K_m(\boldsymbol{x},\boldsymbol{x}).$$

For the second term of (180), we have

$$\begin{aligned}\|\mathcal{I}-P_M\|_2 &= \sup_{f\in\mathcal{B}_{\mathcal{H}_m([0,1]^D)}}\sup_{\boldsymbol{x}\in[0,1]^D}|f(\boldsymbol{x})-P_Mf(\boldsymbol{x})|\\ &\leq \sup_{f\in\mathcal{B}_{\mathcal{H}_m([0,1]^D)}}\sum_{\boldsymbol{\zeta}\in\mathbb{N}^D-[M]^D}|f_{\boldsymbol{\zeta}}|\\ &\leq \left(\sum_{\boldsymbol{\zeta}\in\mathbb{N}^D-[M]^D}(1+\frac{\|\boldsymbol{\zeta}\|_2^2}{m})^{-m}\right)^{\frac{1}{2}},\end{aligned}$$

where the last line is from Hölder inequality and $\forall f\in\mathcal{B}_{\mathcal{H}_m([0,1]^D)}$, $\|f\|_{\mathcal{H}_m([0,1]^D)}\leq 1$. Notice that for $m>D/2$, we have

$$\begin{aligned}\sum_{\boldsymbol{\zeta}\in\mathbb{N}^D-[M]^D}(1+\frac{\|\boldsymbol{\zeta}\|_2^2}{m})^{-m} &= \sum_{\zeta_1\geq M+1}\cdots\sum_{\zeta_D\geq M+1}\left(1+\frac{\sum_{j=1}^D\zeta_j^2}{m}\right)^{-m}\\ &\leq \underbrace{\int_M^\infty\cdots\int_M^\infty}_{D\text{ terms}}(1+\frac{\|\boldsymbol{\zeta}\|_2^2}{m})^{-m}\mathrm{d}\boldsymbol{\zeta}\\ &= \int_0^{2\pi}\cdots\int_0^{2\pi}\int_M^\infty(1+\frac{r^2}{m})^{-m}\det(J(r,\boldsymbol{\theta}))\mathrm{d}r\mathrm{d}\boldsymbol{\theta}\\ &\leq \int_0^{2\pi}\cdots\int_0^{2\pi}\int_M^\infty(1+\frac{r^2}{m})^{-m}r^{D-1}\mathrm{d}r\mathrm{d}\boldsymbol{\theta}\\ &\leq C\frac{1}{2m-D}m^mM^{-2m+D}.\end{aligned}$$

Therefore, we can conclude that $\|\mathcal{I}-P_M\|_2\leq C\frac{1}{2m-D}m^mM^{-2m+D}$ for some C independent of m. Given any $\delta>0$, we can select integer $M=\lceil((2m-D)m^{-m}\delta)^{-\frac{2}{2m-D}}\rceil$ so that

$$\|\mathcal{I}-P_M\|_2\leq\delta,$$

where $\lceil r \rceil$ denotes the ceiling round up of r. Then we can apply Lemma 1 in Kühn (2011) to get

$$\begin{aligned}
H(\delta, \mathcal{B}_{\mathcal{H}_m([0,1]^D)}, \|\cdot\|_{L_\infty([0,1]^D)}) \leq & \operatorname{rank}(P_M) \log(1+\delta^{-1}) \\
\leq & M^D \log(1+\delta^{-1}) \\
\leq & C' \left((2m-D)m^{-m}\delta\right)^{-\frac{2D}{2m-D}} \log(1+\delta^{-1}) \\
= & C'(2m-D)^{-\frac{2D}{2m-D}} m^{\frac{2mD}{2m-D}} \delta^{-\frac{2D}{2m-D}} \log(1+\delta^{-1}),
\end{aligned}$$

for some C' independent of m. ■

Appendix N. Appendix for Detailed Experiments

In this section, we present more details of numerical experiments conducted in Section 5.

Note that in the experiments, our goal is specified by minimizing the l_2 loss in the form of (5). We train the neural network using stochastic gradient descent (SGD) with momentum (0.9), small batch size (10), and learning rate $\beta = 0.01$. We choose a constant weight decay strength (10^{-4}) to focus on the influence of random smoothing in cases with weight decay. We set the number of augmented samples $N = 1000$ and conduct a grid search for the smoothing scale from 0 to 0.6. The simulated data are divided into the training set, validation set, and test set. The validation set is sampled as half the size of the training set, while the size of the test set is fixed at 500. The test results are selected based on the validation set unless otherwise specified and we repeat each experiment 15 times and report the average loss on the test set.

Considering stochastic gradient descent with weight decay, we adopt a candidate list of weight decay strength $\{10^{-3}, 10^{-4}, 10^{-5}\}$. To make a fair comparison, we choose a consistent number of iterations instead of epochs for different training sizes, i.e., given a batch size, the number of epochs gets smaller when the training size becomes larger. Specifically, the number of iterations in cases with weight decay is 10,000. For early stopping without weight decay, we evaluate the validation error every 200 gradient descent steps during training and select the model with the smallest validation error. The maximal step for SGD with early stopping is 100,000. We repeat each experiment 15 times and report the average loss on the test set.

References

Rice (Cammeo and Osmancik). UCI Machine Learning Repository, 2019. DOI: https://doi.org/10.24432/C5MW4Z.

Dry Bean Dataset. UCI Machine Learning Repository, 2020. DOI: https://doi.org/10.24432/C50S4B.

Robert A Adams and John JF Fournier. *Sobolev Spaces*, volume 140. Academic Press, 2003.

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *arXiv preprint arXiv:1901.08584*, 2019a.

Sanjeev Arora, Simon S Du, Zhiyuan Li, Ruslan Salakhutdinov, Ruosong Wang, and Dingli Yu. Harnessing the power of infinitely wide deep nets on small-data tasks. *arXiv preprint arXiv:1910.01663*, 2019b.

Francis R Bach and Michael I Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3(Jul):1–48, 2002.

Yingbin Bai, Erkun Yang, Bo Han, Yanhua Yang, Jiatong Li, Yinian Mao, Gang Niu, and Tongliang Liu. Understanding and improving early stopping for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34:24392–24403, 2021.

Peter L Bartlett and Mikhail Traskin. Adaboost is consistent. *Journal of Machine Learning Research*, 8:2347–2368, 2007.

Moreno Bevilacqua, Tarik Faouzi, Reinhard Furrer, Emilio Porcu, et al. Estimation and prediction using generalized Wendland covariance functions under fixed domain asymptotics. *The Annals of Statistics*, 47(2):828–856, 2019.

Gilles Blanchard and Nicole Krämer. Convergence rates of kernel conjugate gradient for random design regression. *Analysis and Applications*, 14(06):763–794, 2016.

Gilles Blanchard and Nicole Mücke. Optimal rates for regularization of statistical inverse learning problems. *Foundations of Computational Mathematics*, 18(4):971–1013, 2018.

Avrim Blum, Travis Dick, Naren Manoj, and Hongyang Zhang. Random smoothing might be unable to certify ℓ_∞ robustness for high-dimensional images. *Journal of Machine Learning Research*, 21(211):1–21, 2020. URL `http://jmlr.org/papers/v21/20-209.html`.

Peter Bühlmann and Bin Yu. Boosting with the l_2-loss: Regression and classification. *Journal of American Statistical Association*, 98:324–340, 2002.

Hans-Joachim Bungartz and Michael Griebel. A note on the complexity of solving Poisson's equation for spaces of bounded mixed derivatives. *Journal of Complexity*, 15(2):167–199, 1999.

Yuan Cao and Quanquan Gu. Generalization error bounds of gradient descent for learning over-parameterized deep relu networks. In *AAAI Conference on Artificial Intelligence*, volume 34, pages 3349–3356, 2020.

Andrea Caponnetto and Yuan Yao. Cross-validation based adaptation for regularization operators in learning theory. *Analysis and Applications*, 8(2):161–183, 2010. CBCL Paper #265/AI Technical Report #063, Massachusetts Institute of Technology, Cambridge, MA, September, 2006.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Ieee Symposium on Security and Privacy (sp)*, pages 39–57. Ieee, 2017.

Daniel Cervone and Natesh S Pillai. Gaussian process regression with location errors. *arXiv preprint arXiv:1506.08256*, 2015.

Haoyuan Chen, Liang Ding, and Rui Tuo. Kernel packet: An exact and scalable algorithm for gaussian process regression with matérn correlations. *Journal of Machine Learning Research*, 23(127):1–32, 2022.

Lin Chen and Sheng Xu. Deep neural tangent kernel and laplace kernel have the same rkhs. *arXiv preprint arXiv:2009.10683*, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.

Andrew Chernih and Simon Hubbert. Closed form representations and properties of the generalised Wendland functions. *Journal of Approximation Theory*, 177:17–33, 2014.

İlkay Çinar, Murat Koklu, and Şakir Taşdemir. Raisin. UCI Machine Learning Repository, 2023. DOI: https://doi.org/10.24432/C5660T.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pages 1310–1320. PMLR, 2019.

Noel Cressie and John Kornak. Spatial statistics in the presence of location error with an application to remote sensing of the environment. *Statistical Science*, 18(4):436–456, 2003.

SÁNDOR Csörgő. Rates of uniform convergence for the empirical characteristic function. *Acta Sci. Math.(Szeged)*, 48(1–4):97–102, 1985.

Ronald A DeVore and Robert C Sharpley. Besov spaces on domains in R^d. *Transactions of the American Mathematical Society*, 335(2):843–864, 1993.

Lee H Dicker, Dean P Foster, Daniel Hsu, et al. Kernel ridge vs. principal component regression: Minimax bounds and the qualification of regularization operators. *Electronic Journal of Statistics*, 11(1):1022–1047, 2017.

Aymeric Dieuleveut and Francis Bach. Nonparametric stochastic approximation with large step-sizes. *The Annals of Statistics*, 44(4):1363–1399, 2016.

Liang Ding and Xiaowei Zhang. Sample and computationally efficient stochastic kriging in high dimensions. *Operations Research*, 2022.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.

Dinh Dũng. Deep ReLU neural networks in high-dimensional approximation. *Neural Networks*, 142:619–635, 2021.

Dinh Dung, Vladimir Temlyakov, and Tino Ullrich. *Hyperbolic Cross Approximation*. Springer, 2018.

Mona Eberts and Ingo Steinwart. Optimal regression rates for svms using gaussian kernels. *Electronic Journal of Statistics*, 7:1–42, 2013.

David Eric Edmunds and Hans Triebel. *Function Spaces, Entropy Numbers, Differential Operators*, volume 120. Cambridge University Press, 2008.

H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Kluwer Academic Publishers, 1996.

Lawrence C Evans. Partial differential equations (graduate studies in mathematics, vol. 19). *Instructor*, 67, 2009.

Gregory E Fasshauer and Michael J McCourt. *Kernel-based approximation methods using MATLAB*, volume 19. World Scientific Publishing Company, 2015.

R. A. Fisher. Iris. UCI Machine Learning Repository, 1988. DOI: https://doi.org/10.24432/C56C76.

Zhidong Gao, Rui Hu, and Yanmin Gong. Certified robustness of graph classification against topology attack with randomized smoothing. In *GLOBECOM 2020-2020 IEEE Global Communications Conference*, pages 1–6. IEEE, 2020.

Jochen Garcke, Michael Griebel, and Michael Thess. Data mining with sparse grids. *Computing*, 67(3):225–253, 2001.

Amnon Geifman, Abhay Yadav, Yoni Kasten, Meirav Galun, David Jacobs, and Basri Ronen. On the similarity between the laplace and neural tangent kernels. *Advances in Neural Information Processing Systems*, 33:1451–1461, 2020.

Amnon Geifman, Meirav Galun, David Jacobs, and Basri Ronen. On the spectral bias of convolutional neural tangent and gaussian process kernels. *Advances in Neural Information Processing Systems*, 35:11253–11265, 2022.

T Gneiting. Stationary covariance functions for space-time data. *Journal of the American Statistical Association*, 97:590–600, 2002.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep Learning*, volume 1. MIT Press, 2016.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Yves Grandvalet, Stéphane Canu, and Stéphane Boucheron. Noise injection: Theoretical prospects. *Neural Computation*, 9(5):1093–1108, 1997.

Arthur Gretton. A simpler condition for consistency of a kernel independence test. Technical report, University College London, 2015.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International Conference on Algorithmic Learning Theory*, pages 63–77. Springer, 2005.

Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. A kernel statistical test of independence. *Advances in Neural Information Processing Systems*, 20, 2007.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

Zheng-Chu Guo, Shao-Bo Lin, and Ding-Xuan Zhou. Learning theory of distributed spectral algorithms. *Inverse Problems*, 33(7):074009, 2017.

Thomas Hamm and Ingo Steinwart. Adaptive learning rates for support vector machines working on data with low intrinsic dimension. *The Annals of Statistics*, 49(6):3153–3180, 2021a.

Thomas Hamm and Ingo Steinwart. Intrinsic dimension adaptive partitioning for kernel methods. *arXiv preprint arXiv:2107.07750*, 2021b.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234. PMLR, 2016.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2001.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021.

Tianyang Hu, Wenjia Wang, Cong Lin, and Guang Cheng. Regularization matters: A non-parametric perspective on overparametrized neural network. In *International Conference on Artificial Intelligence and Statistics*, pages 829–837. PMLR, 2021.

Tianyang Hu, Jun Wang, Wenjia Wang, and Zhenguo Li. Understanding square loss in training overparametrized neural network classifiers. *Advances in Neural Information Processing Systems*, 35:16495–16508, 2022.

Kaixuan Huang, Yuqing Wang, Molei Tao, and Tuo Zhao. Why do deep residual networks generalize better than deep feedforward networks?—a neural tangent kernel perspective. *Advances in Neural Information Processing Systems*, 33:2698–2709, 2020.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, pages 8571–8580, 2018.

Ryuichi Kanoh and Mahito Sugiyama. A neural tangent kernel perspective of infinite tree ensembles. *arXiv preprint arXiv:2109.04983*, 2021.

Samuel Kotz, Tomasz Kozubowski, and Krzysztof Podgórski. *The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance*. Number 183. Springer Science & Business Media, 2001.

Tomasz J Kozubowski, Krzysztof Podgórski, and Igor Rychlik. Multivariate generalized laplace distribution and related random fields. *Journal of Multivariate Analysis*, 113: 59–72, 2013.

Anders Krogh and John A Hertz. A simple weight decay can improve generalization. In *Advances in Neural Information Processing Systems*, pages 950–957, 1992.

Thomas Kühn. Covering numbers of gaussian reproducing kernel hilbert spaces. *Journal of Complexity*, 27(5):489–499, 2011.

Thomas Kühn, Winfried Sickel, and Tino Ullrich. Approximation of mixed order Sobolev functions on the d-torus: Asymptotics, preasymptotics, and d-dependence. *Constructive Approximation*, 42(3):353–398, 2015.

Ronald Rapha L Coifman and Guido L Weiss. *Transference Methods in Analysis*, volume 31. American Mathematical Soc., 1977.

Giovanni Leoni. *A first course in Sobolev spaces*. American Mathematical Soc., 2017.

Mingchen Li, Mahdi Soltanolkotabi, and Samet Oymak. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 4313–4324. PMLR, 2020.

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, pages 8157–8166, 2018.

Zhiyuan Li, Ruosong Wang, Dingli Yu, Simon S Du, Wei Hu, Ruslan Salakhutdinov, and Sanjeev Arora. Enhanced convolutional neural tangent kernels. *arXiv preprint arXiv:1911.00809*, 2019.

Junhong Lin and Lorenzo Rosasco. Optimal rates for multi-pass stochastic gradient methods. *Journal of Machine Learning Research*, 18(1):3375–3421, 2017.

Junhong Lin, Lorenzo Rosasco, and Ding-Xuan Zhou. Iterative regularization for learning with convex loss functions. *Journal of Machine Learning Research*, 17(1):2718–2755, 2016.

Shao-Bo Lin, Xin Guo, and Ding-Xuan Zhou. Distributed learning with regularized least squares. *Journal of Machine Learning Research*, 18(1):3202–3232, 2017.

Pascal Massart. *Concentration Inequalities and Model Selection*, volume 6. Springer, 2007.

Akshay Mehra, Bhavya Kailkhura, Pin-Yu Chen, and Jihun Hamm. How robust are randomized smoothing based defenses to data poisoning? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13244–13253, 2021.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *International Conference on Machine Learning*, pages 807–814, 2010.

Vern I Paulsen and Mrinal Raghupathi. *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*, volume 152. Cambridge University Press, 2016.

Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. *Advances in Neural Information Processing Systems*, 31, 2018.

Lutz Prechelt. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer, 1998.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *Journal of Machine Learning Research*, 15(1):335–366, 2014.

Christian Rieger and Holger Wendland. Sampling inequalities for sparse grids. *Numerische Mathematik*, 136:439–466, 2017.

Elan Rosenfeld, Ezra Winston, Pradeep Ravikumar, and Zico Kolter. Certified robustness to label-flipping attacks via randomized smoothing. In *International Conference on Machine Learning*, pages 8230–8241. PMLR, 2020.

Yunus Saatçi. *Scalable inference for structured Gaussian process models*. PhD thesis, University of Cambridge, 2012.

Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. *Advances in Neural Information Processing Systems*, 32, 2019.

Clayton Scott and Robert D Nowak. Minimax-optimal classification with dyadic decision trees. *IEEE Transactions on Information Theory*, 52(4):1335–1353, 2006.

Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48, 2019.

Elias M Stein. *Singular integrals and differentiability properties of functions*, volume 2. Princeton University Press, 1970.

Ingo Steinwart, Don R Hush, and Clint Scovel. Optimal rates for regularized least squares regression. In *COLT*, pages 79–93, 2009.

Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, pages 1040–1053, 1982.

Zoltán Szabó and Bharath K Sriperumbudur. Characteristic and universal tensor product kernels. *J. Mach. Learn. Res.*, 18:233–1, 2017.

Rui Tuo and C. F. Jeff Wu. A theoretical framework for calibration in computer models: Parametrization, estimation and convergence properties. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):767–795, 2016.

Rui Tuo, Yan Wang, and CF Wu. On the improved rates of convergence for Matérn-type kernel ridge regression, with application to calibration of computer models. *arXiv preprint arXiv:2001.00152*, 2020.

Sara van de Geer. *Empirical Processes in M-estimation*. Cambridge University Press, 2000.

Richard S Varga. *Gershgorin and His Circles*, volume 36. Springer Science & Business Media, 2010.

Grace Wahba. *Spline Models for Observational Data*, volume 59. SIAM, 1990.

Binghui Wang, Xiaoyu Cao, Neil Zhenqiang Gong, et al. On certifying robustness against backdoor attacks via randomized smoothing. *arXiv preprint arXiv:2002.11750*, 2020.

Wenjia Wang. On the inference of applying Gaussian process modeling to a deterministic function. *Electronic Journal of Statistics*, 15(2):5014–5066, 2021.

Wenjia Wang and Bing-Yi Jing. Gaussian process regression: Optimality, robustness, and relationship with kernel ridge regression. *Journal of Machine Learning Research*, 23(193): 1–67, 2022.

Wenjia Wang, Xiaowei Yue, Benjamin Haaland, and CF Jeff Wu. Gaussian processes with input location error and applications to the composite parts assembly process. *SIAM/ASA Journal on Uncertainty Quantification*, 10(2):619–650, 2022.

Yuting Wei, Fanny Yang, and Martin J Wainwright. Early stopping for kernel boosting algorithms: A general analysis with localized complexities. *Advances in Neural Information Processing Systems*, 30, 2017.

Holger Wendland. *Scattered Data Approximation*, volume 17. Cambridge University Press, 2004.

Christopher KI Williams and Carl Edward Rasmussen. *Gaussian Processes for Machine Learning*, volume 2. MIT Press Cambridge, MA, 2006.

Andrew Wilson and Hannes Nickisch. Kernel interpolation for scalable structured gaussian processes (kiss-gp). In *International Conference on Machine Learning*, pages 1775–1784. PMLR, 2015.

Qiang Wu, Yiming Ying, and Ding-Xuan Zhou. Learning rates of least-square regularized regression. *Foundations of Computational Mathematics*, 6(2):171–192, 2006.

Yun Yang and David B Dunson. Bayesian manifold regression. *The Annals of Statistics*, 44(2):876–905, 2016.

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

Gui-Bo Ye and Ding-Xuan Zhou. Learning and approximation by gaussians on riemannian manifolds. *Advances in Computational Mathematics*, 29(3):291–310, 2008.

Gui-Bo Ye and Ding-Xuan Zhou. Svm learning and lp approximation by gaussians on riemannian manifolds. *Analysis and Applications*, 7(03):309–339, 2009.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Tong Zhang and Bin Yu. Boosting with early stopping: Convergence and consistency. *The Annals of Statistics*, 33(4):1538–1579, 2005.

Shuang Zhou, Debdeep Pati, Tianying Wang, Yun Yang, and Raymond J Carroll. Gaussian processes with errors in variables: Theory and computation. *arXiv preprint arXiv:1910.06235*, 2019.