Random Smoothing Regularization in Kernel Gradient Descent

December 2024

Abstract

Machine learning has emerged as a rapidly growing field with wide-ranging applications. While the practical successes of machine learning are widely recognized, the deep mathematical foundations underlying these methods are often overlooked and attributed solely to statistics and theoretical computer science. This expository paper seeks to lay clear the mathematics in recent machine learning research. In particular, I examine Ding et al., 2023; which presents a framework for analyzing random smoothing regularization as convolution-based smoothing kernels attaining optimal convergence rates. This work builds the necessary mathematical background to understand these results, then formulates them step-by-step, assuming only an undergraduate understanding of machine learning, Fourier analysis, function spaces, and measure theory.

1 Introduction

Random smoothing is a data augmentation technique used in machine learning to improve the generalization and robustness of machine learning models. With regards to the input data, generalization encourages models to learn meaningful patterns rather than memorizing a training set, whereas robustness allows models to remain resistant to variations or noise. Random smoothing aims to transform a base model into a smoothed one that is provably robust to perturbations. By adding random noise, typically Gaussian or Laplace, to the input data during the training process, one can simulate naturally occurring variations in real-world data.

In particular, random smoothing can be considered a form of regularization, a class of techniques which address model overfitting through additional constraints or information in the learning process. Common examples include L^1 (ridge) and L^2 (LASSO) regularization, which add penalty terms to the loss function, named for their usage of L^p norms. Random smoothing regularization is implicit, as it does not explicitly perform subset selection, modify parameters, or alter loss function.

However, as Ding et al.(2023) remarks, there is a lack of literature on the regularizing effects of random smoothing. Their work introduces several theoretical findings for the classical Sobolev spaces. Firstly, for Sobolev spaces of low intrinsic dimensionality, they present optimal convergence rates for polynomial and Gaussian random smoothing is proven. Secondly, for mixed smooth Sobolev spaces with a tensor structure, they show an optimal convergence rate for polynomial random smoothing. For the sake of length, I will cover only the first part of the paper.

Section 2 introduces concepts necessary for the problem setting including nonparametric regression, Sobolev spaces, Reproducing Kernel Hilbert Spaces (RKHS), random smoothing, and convolution.

Section 3 introduces the problem setting analogously to Section 3 of Ding et al.(2023), with elaboration on the assumptions.

Section 4 presents a step-by-step approach to understanding the main results on Ding et al.(2023).

2 Background

2.1 Nonparametric Regression

In many machine learning problems, we observe pairs of data (x_i, y_i) where $x_i \in \mathbb{R}^p$ is the feature vector and $y_i \in \mathbb{R}$ is the response. In parametric regression, we approximate the true function by finding functions constrained by a certain number of parameters (i.e., linear regression will optimize over a space of possible linear functions).

On the other hand, nonparametric regression assumes that

$$y_i = f^*(x_i) + \epsilon_i,$$

with $f*$ a smooth function of x and ϵ_i i.i.d. noise variables with zero mean and finite variance. In particular, in this problem, x i.i.d. following marginal distribution P_X with support

$$\text{supp}(P_X) = \Omega \subset \mathbb{R}^D.$$

As with any regression problem, we seek to recover the function $f*$ to model the observed data. Kernel methods are a popular way to solve such regression problems.

2.2 Kernel Methods and RKHS

A kernel is a bivariate function $K : \Omega \times \Omega \mapsto \mathbb{R}$. In particular, we are interested in Mercer kernels, which, roughly speaking, optimize the fit of a function to data subject to regularization or penalization. Mercer kernels share the property that, for any set of points $x_1, ..., x_n$, the $n \times n$ matrix $K = [K(x_i, x_j)]$, also called the gram matrix, is positive semidefinite.

We can define a Hilbert space from the kernel, which gives us an infinite dimensional space of functions with a geometry. Its inner product defines such a generalization through a distance function and angle.

To understand the construction of this kernel-defined Hilbert space, we first create a set of basis functions based on K. Fix z and think of $K(z, x)$ as a function of x, i.e.,

$$K(z, x) = K_z(x)$$

is a function of the second argument, where we have fixed the first argument. By the kernel's positive semidefinite property, we can define an inner product and norm over the span of these basis functions:

$$\begin{aligned}
f(x) &= \sum_r \alpha_r K_{z_r}(x) && (1)\\
g(x) &= \sum_s \beta_s K_{y_s}(x) && (2)\\
\langle f, g \rangle_K &= \sum_r \sum_s \alpha_r \beta_s K(z_r, y_s) && (3)\\
&= \alpha^T \mathbb{K} \beta && (4)
\end{aligned}$$

where $K = [K(z_r, y_s)]$. Note this implies $||f||_K^2 = \alpha^T \mathbb{K} \alpha \geq 0$. This defines the linear space

$$F_K(\Omega) = \left\{ \sum_{k=1}^{n} \beta_k K(\cdot, x_k) : \beta_k \in \mathbb{R}, x_k \in \Omega, n \in \mathbb{N} \right\}.$$

With the bilinear form defined in (3), the reproducing kernel Hilbert space (abbreviated RKHS) $\mathcal{H}_K(\Omega)$ generated by the kernel K is defined as the closure of $F_K(\Omega)$ under the inner product $\langle \cdot, \cdot \rangle_K$. The norm of this RKHS is given as

$$||f||_{\mathcal{H}_K(\Omega)} = \sqrt{\langle f, f \rangle_{\mathcal{H}_K(\Omega)}},$$

with $\langle \cdot, \cdot \rangle_{\mathcal{H}_K(\Omega)}$ induced by $\langle \cdot, \cdot \rangle_K$.

It is called a reproducing kernel Hilbert space because

$$\langle f, K_x(\cdot) \rangle_K = f(x),$$

i.e., the kernel reproduces the values of the functions through the inner products.

RKHS provide the basis for kernel methods in machine learning, which this problem will build off of.

2.2.1 Random Smoothing Kernel Function

We can define an empirical random smoothing kernel function K_S corresponding to an original kernel K by

$$K_S(x_l - x_j) \overset{\text{def}}{=} \frac{1}{N^2} \sum_{k_1=1}^{N} \sum_{k_2=1}^{N} K(x_l + \epsilon_{k_1} - (x_j + \epsilon_{k_2})).$$

The expectation of K_S with respect to the noise ϵ_k is the convolution of $K * p_\epsilon$, defined

$$(K * p_\epsilon)(s) = \int K(t) p_\epsilon(s-t) dt,$$

following the general definition of the convolution operator. This convoluted kernel function $K * p_\epsilon$ is called the random smoothing kernel function. We show later the importance of this interpretation of the random smoothing kernel function for analyzing convergence rates.

2.3 Sobolev Spaces

A Sobolev space, denoted by $W^{k,p}(\Omega)$, is a vector space of functions defined on an open subset Ω of $\mathbb{R}^n$. It consists of functions u that belong to the $L^p(\Omega)$ space and whose weak derivatives up to order k also belong to $L^p(\Omega)$.

- **Weak Derivatives:** Instead of requiring the classical differentiability of functions, in a Sobolev space we can require only weak derivatives. A function u has a weak derivative $D^\alpha u$ if there exists a function $v \in L^p(\Omega)$ such that for all test functions $\phi \in C_0^\infty(\Omega)$, the following holds:

$$\int_\Omega v\phi \, dx = (-1)^{|\alpha|} \int_\Omega u D^\alpha \phi \, dx$$

  Here, D^α denotes the mixed partial derivative corresponding to the multi-index α.

- **L^p-Integrability:** For a function u to be in $W^{k,p}(\Omega)$, both u and its weak derivatives up to order k must be L^p-integrable.

- **Norms:** The norm in a Sobolev space $W^{k,p}(\Omega)$ is defined as:

$$\|u\|_{W^{k,p}(\Omega)} = \left(\sum_{|\alpha| \leq k} \|D^\alpha u\|_{L^p(\Omega)}^p \right)^{1/p}$$

  for $1 \leq p < \infty$, and for $p = \infty$, it is defined as the maximum of the L^∞ norms of u and its derivatives up to order k.

We can understand Sobolev spaces as important to the problem setting due to the effects of random smoothing. As an implicit form of regularization, we can interpret random smoothing as encouraging a learned function to be smoother and therefore more likely to fall into a specific Sobolev space.

2.4 Covering Number

One of the assumptions in the paper regards the intrinsic dimension of Ω, which is determined by the covering number and can be thought of as a measure of how complex the region is. Roughly speaking, we can understand the covering number as the minimum number of "balls" of a given radius needed to fully cover a dataset, which gives a representation of its complexity.

For a subset $\mathcal{A} \subset \mathcal{G}$ where $\mathcal{G}$ a normed space and some fixed $\delta > 0$, the **covering number** of $\mathcal{A}$, denoted $\mathcal{N}_{\mathcal{G}}(\delta, \mathcal{A})$, is the smallest integer M such that $\mathcal{A}$ can be covered by M balls of radius δ and centers $x_1, ..., x_M \in \mathcal{G}$.

2.5 Lipschitz Domain

A domain with Lipschitz boundary, also known as a Lipschitz domain, is defined as follows.

Let $n \in \mathbb{N}$. Let D be a domain of $\mathbb{R}^n$ and let ∂D denote the boundary of D. Then D is called a **Lipschitz domain** if for every point $p \in \partial D$ there exists a hyperplane H of dimension $n-1$ through p, a Lipschitz-continuous function $g : H \to \mathbb{R}$ over that hyperplane, and reals $r > 0$ and $h > 0$ such that

- $D \cap C = \{x + y\vec{n} \mid x \in B_r(p) \cap H, -h < y < g(x)\}$
- $(\partial D) \cap C = \{x + y\vec{n} \mid x \in B_r(p) \cap H, g(x) = y\}$

where

- $\vec{n}$ is a unit vector that is normal to H,
- $B_r(p) := \{x \in \mathbb{R}^n \mid \|x - p\| < r\}$ is the open ball of radius r,
- $C := \{x + y\vec{n} \mid x \in B_r(p) \cap H, -h < y < h\}$.

In other words, at each point of its boundary, D is locally the set of points located above the graph of some Lipschitz function. Why is this interpretation true? Consider the case where $n = 2$. Then the hyperplane H in $\mathbb{R}^2$ would simply be a line. Since it goes through a point in the boundary of D, we can think of this line as tangent to ∂D. Imagining ourselves in the $x - y$ plane for familiarity, notice the x axis is, as a line, a hyperplane representing the domain of the function. So we can imagine H as the "x axis" of g, whereas H^c, which is spanned by $\vec{n}$, as the "y axis." Then, locally, the graph of g would be given by vectors $x + g(x)\vec{n}$ for $x \in H$.

Lipschitz boundaries are an important restriction that essentially ensure that the boundary of a domain is smooth or regular enough to guarantee well-defined solutions, especially for problems involving derivatives. Since this paper works with Sobolev spaces, we will later see an assumption of Lipschitz boundaries.

2.6 Representer Theorem

The Representer Theorem is commonly used in kernel problems and appears in the proofs of the main results. Essentially, the theorem says that the minimizer of a regularized risk function on an RKHS can be represented as a finite linear combination of kernel products.

Formally, let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive definite kernel on a non-empty set $\mathcal{X}$ with a corresponding RKHS $\mathcal{H}_k$. Suppose we have

- a training sample $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathbb{R}$,
- a strictly increasing real-valued function $g : [0, \infty) \to \mathbb{R}$,
- an arbitrary error function $E : (\mathcal{X} \times \mathbb{R})^n \to \mathbb{R} \cup \{\infty\}$.

Then we can define a regularized empirical risk functional on $\mathcal{H}_k$:

$$f \mapsto E\big((x_1, y_1, f(x_1)), \ldots, (x_n, y_n, f(x_n))\big) + g(\|f\|).$$

Where any minimizer, f^* of the empirical risk

$$f^* = \arg\min_{f \in \mathcal{H}_k} \{E((x_1, y_1, f(x_1)), \ldots, (x_n, y_n, f(x_n))) + g(\|f\|)\}, \quad (*)$$

admits a representation of the form:

$$f^*(\cdot) = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i),$$

where $\alpha_i \in \mathbb{R}$ for all $1 \leq i \leq n$.

3 Problem Setting

We examine the nonparametric regression problem setting (2.1). The paper examines two cases; firstly, where the function space $\mathcal{H}(\Omega)$ is a Sobolev space $\mathcal{W}^m(\Omega)$ and the data has low intrinsic dimension. The second case is where the function space is a tensor Sobolev space. The observed data (x_i, y_i) are assumed x i.i.d. following marginal distribution P_X uniform.

We use RKHS to recover f^*. Using the linear space $F_K(\Omega)$ and the RKHS $\mathcal{H}_K(\Omega)$ defined in (2.2), we can characterize the RKHS for stationary kernel K (e.g., $K(x - x') = K(x, x')$) via the Fourier transform, $\mathcal{F}$: for a function $g \in L_1(\mathbb{R}^D)$,

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

Using the following theorem, we can connect Sobolev spaces to the RKHS.

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

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

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

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

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

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

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

Only Sobolev spaces with $m > D/2$ are considered in the paper. This is due to a result of the Sobolev Embedding Theorem, which tells under which conditions functions in a Sobolev space $W^{m,p}(\Omega)$ will also possess additional regularity, such as continuity, differentiability, or even higher regularity.

A critical case of the Sobolev Embedding Theorem is when the space $W^{m,p}(\Omega)$ embeds into the space of continuous functions $C(\Omega)$. Specifically, the theorem states that if $m > \frac{D}{p}$, then $W^{m,p}(\Omega)$

embeds into the space of continuous functions. Hence if f $m > \frac{D}{2}$ and $p = 2$ (which corresponds to the L^2 setting), the Sobolev space $W^{m,2}(\Omega)$ will embed into the space of continuous functions.

Then, comparing Theorem 1 and (1), if

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

for some two constants $c_1, c_2 > 0$, we see $\mathcal{W}^m(\mathbb{R}^D)$ coincides with the reproducing kernel Hilbert space $\mathcal{H}_K(\mathbb{R}^D)$ with equivalent norms. This connects the Sobolev spaces examined in the paper to RKHS.

3.1 Assumptions

The results of Ding et al.(2023) occur within several crucial assumptions. We lay out each of these assumptions, numbered according to the original paper, below.

3.1.1 Main Assumptions

These assumptions are used for all results.

Assumption 1 *The error ϵ_j's in the nonparametric regression problem setting (discussed in 2.1) are i.i.d. sub-Gaussian, as defined in 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'.

This sub-Gaussian assumption is important because it provides tighter control over the tails of the distribution, rather than just assuming finite variance. Since sub-Gaussian variables have exponentially decaying tails, they can be useful for deriving concentration inequalities and allow for stronger theoretical guarantees.

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

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

for some positive constants c_1 and c_2.

The important of this assumption was discussed earlier in the problem setting section (3), ensuring that the Sobolev space under consideration embeds into the space of continuous functions and is connected by the Fourier transform to a stationary RKHS.

Assumption 3 *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.

The kernel function is given to have a tensor structure such that the Fourier transform of each component has an algebraic decay. This, like Assumption 2, is because we are working with a sufficiently smooth Sobolev space which coincides with the RKHS.

Assumption 4 *The elements of ε_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.*

1. *(Polynomial noise) There exists $m_\varepsilon > D/2$ such that the characteristic function of ε_k satisfies*
$$c_1(1+\sigma_n^2\|\omega\|_2^2)^{-m_\varepsilon} \le \mathbb{E}(e^{i\omega^T\varepsilon_k}) \le c_2(1+\sigma_n^2\|\omega\|_2^2)^{-m_\varepsilon}, \forall \omega \in \mathbb{R}^D.$$

2. *(Tensor Polynomial noise) There exists $m_\varepsilon > 1/2$ such that the characteristic function of ε_k satisfies*
$$c_1\prod_{j=1}^{D}(1+\sigma_n^2\omega_j^2)^{-m_\varepsilon} \le \mathbb{E}(e^{i\omega^T\varepsilon_k}) \le c_2\prod_{j=1}^{D}(1+\sigma_n^2\omega_j^2)^{-m_\varepsilon}, \forall \omega = (\omega_1,\ldots,\omega_D) \in \mathbb{R}^D.$$

3. *(Gaussian noise) The elements of ϵ_k are normally distributed with variance σ_n^2.*

The constants c_1 and c_2 are not dependent on σ_n and m_ϵ. σ_n is referred to as the smoothing scale.

Assumption 4 gives conditions on the noise terms (ϵ_k's) which consider the 3 problem settings of the paper: polynomial, tensor, and Gaussian smoothing. Hence each of the assumptions is simply the restriction on the noise terms such that the problem falls into a certain category (e.g., requiring the noise to be normally distributed with some squared variance is a Gaussian problem).

Assumption 5 *There exist positive constants c_1 and $d \le D$ such that for all $\delta \in (0,1)$, we have*
$$\mathcal{N}_{\ell D_\infty}(\delta,\Omega) \le c_1\delta^{-d},$$
where ℓD_∞ is the $\mathbb{R}^D$ space with the L^∞ norm and $\mathcal{N}$ is the covering number defined in (2.4).

This is an assumption of low intrinsic dimension; that is, the covering number of the given space for balls of radius $\delta \in (0,1)$ is bounded by some positive constants. In particular, the covering number is proportional to δ^{-d}, where d is less than the dimension D of the $\mathbb{R}^D$ space.

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 $m_f = argsup_{m>D/2}\{m : f^*\mathcal{W}^m(\Omega_1)\}$, and $m_f > D/2$.*

The importance of the Lipschitz boundary is explained in (2.5). Essentially, this assumption supposes that the true function being approximated is well-defined in a Sobolev space, with the $D/2$ term appearing for reasons previously discussed in (3).

4 Results

We are now able to discuss the main results of this paper. We begin with polynomial smoothing.

Theorem 2 (Polynomial smoothing)**.** *Suppose Assumptions 1, 2, 4 (C1), 5 and 6 are satisfied. Let* $f_1(\mathbf{x})$ *be as in (12) and* $\beta = n^{-1}C_1$ *with the positive constant* $C_1 \le (2\sup_{\mathbf{x}\in\mathbb{R}^D} K_S(\mathbf{x}))^{-1}$*. Suppose the smoothing scale* $\sigma_n \asymp n^{\nu}$ *with* $\nu \le 0$*. 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* $t \ge C_2(\frac{m_f}{2m_f+d} + 1/2)\log n/(\log(1-\alpha))$ *for some positive constant* C_2*.*

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

$$\|f_t - f^*\|_{L_2(P_X)}^2 = \mathcal{O}_{\mathbb{P}}\left(n^{-\frac{2m_f}{2m_f+d}}(\log n)^{2m_f+1}\right)$$

for $N > N_0$*, where* N *is the number of augmentations, and* N_0 *depends on* n *and the iteration number* t*.*

Several lemmas are used in this proof. I omit their proofs, as they are included in the appendices of the original paper, and their inclusion would add around 30 pages to this work without clarifying the main results. Firstly,

Lemma 1. *Suppose the conditions of Theorem 8 are fulfilled. Let* f_n^* *be the solution to the optimization problem*

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

Then if $m_0 \le m_f$*, we have*

$$\|f^* - f_n^*\|_{L_2(P_X)}^2 + \lambda_n \|f_n^*\|_{\mathcal{H}_{\bar{K}_S}(\Omega)}^2 \le C_1 max\left((\lambda_n(m_\varepsilon+1)^{m_\varepsilon}\sigma_n^{2m_\varepsilon})^{\frac{m_f}{m_0+m_\varepsilon}}, \lambda_n\right).$$

and if $m_0 > m_f$, we have

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

Here the constants C_1 and C_2 are independent with m_ε.

Lemma 2. *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*

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

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

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

where $y = (y_1, ..., y_n)^T$. Suppose

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

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

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

Then we have

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

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

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

then

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

Lemma 3. *(Lemma F.5 of Wang, 2021) Assume for class $\mathcal{G}$, $\sup_{g\in\mathcal{G}}\|g\|_{L_\infty(\Omega)} \leq c < 1$, and the bracket entropy $H_B(\delta_n, \mathcal{G}, \|\cdot\|_{L_2(P_X)}) \leq \frac{n\delta_n^2}{1200c^2}$, and $n\delta_n^2 \to \infty$, where $0 < \delta_n < 1$. Then we have*

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

and

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

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

Lemma 4. *(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*

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

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

Lemma 5. *(Error of Data Augmentation) Suppose Assumption 2 or 3, and Assumption 4 are satisfied. Furthermore, assume that*

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

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

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

where the probability is with respect to the augmentation ε.

We now proceed to proving the first theorem. Firstly, in the case of no weight decay:

Proof. **Without Weight Decay.**

By the triangle inequality, it can be seen that

$$\|f_t - f^*\|_{L_2(P_X)} \leq \|f_t - g_t\|_{L_2(P_X)} + \|g_t - f^*\|_{L_2(P_X)}, \tag{1}$$

where g_t is as in the analysis on gradient update (below in 4.2).

By Lemma 5, the first term $\|f_t - g_t\|_{L_2(P_X)}$ in (1) can be bounded by

$$\|f_t - g_t\|_{L_2(P_X)} \leq C_{10}\|f_t - g_t\|_{L_\infty(\Omega)} = \mathcal{O}_{\mathbb{P}}\left(\frac{n^2\sqrt{\log N/N}}{\eta_n(\tilde{\mathbf{K}})^2}\right), \tag{2}$$

as long as

$$\frac{1}{2}\eta_n(\tilde{\mathbf{K}}) \geq n\sqrt{\frac{\log N}{N}}. \tag{3}$$

Choose

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

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

$$\|f_t - g_t\|_{L_2(P_X)} = \mathcal{O}_{\mathbb{P}}(n^{-1/2}). \tag{5}$$

$$J_2 = \|g_t - f^*\|_n^2 = \frac{1}{n}\|g_t(\mathbf{X}) - f^*(\mathbf{X})\|_2^2. \tag{6}$$

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

$$\bar{g} = \underset{f\in\mathcal{H}_{\bar{K}_S}(\Omega)}{\operatorname{argmin}} \|f - \mathbf{y}\|_n^2 + \lambda_n\|f\|_{\mathcal{H}_{\bar{K}_S}(\Omega)}^2. \tag{7}$$

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

$$\bar{g}(\mathbf{X}) - f^*(\mathbf{X}) = n\lambda_n(\tilde{\mathbf{K}} + n\lambda_n\mathbf{I})^{-1}f^*(\mathbf{X}) + \tilde{\mathbf{K}}(\tilde{\mathbf{K}} + n\lambda_n\mathbf{I})^{-1}\boldsymbol{\epsilon} = \mathbf{q}_1 + \mathbf{q}_2. \tag{8}$$

Since

$$g_t(\mathbf{X}) = \left(\mathbf{I} - (\mathbf{I} - \beta\tilde{\mathbf{K}})^t\right)\mathbf{y}, \tag{9}$$

we have that

$$g_t(\mathbf{X}) - f^*(\mathbf{X}) = -(\mathbf{I} - \beta\tilde{\mathbf{K}})^t f^*(\mathbf{X}) + \left(\mathbf{I} - (\mathbf{I} - \beta\tilde{\mathbf{K}})^t\right) \boldsymbol{\epsilon}. \tag{10}$$

By the Cauchy-Schwarz inequality, (6), and (10), it then follows that

$$\begin{aligned} nJ_2 &\leq 2(f^*(\mathbf{X}))^T(\mathbf{I} - \beta\tilde{\mathbf{K}})^{2t} f^*(\mathbf{X}) + 2\boldsymbol{\epsilon}^T(\mathbf{I} - (\mathbf{I} - \beta\tilde{\mathbf{K}})^t)^2 \boldsymbol{\epsilon} && (11)\\ &= 2nJ_{21} + 2nJ_{22}, && (12)\end{aligned}$$

and

$$\begin{aligned} n\|\bar{g} - f^*\|_n^2 &\leq 2(n\lambda_n)^2(f^*(\mathbf{X}))^T(\tilde{\mathbf{K}} + n\lambda_n\mathbf{I})^{-2} f^*(\mathbf{X}) + 2\boldsymbol{\epsilon}^T(\tilde{\mathbf{K}} + n\lambda_n\mathbf{I})^{-1}\tilde{\mathbf{K}}^2(\tilde{\mathbf{K}} + n\lambda_n\mathbf{I})^{-1}\boldsymbol{\epsilon} && (13)\\ &= 2\|\mathbf{q}_1\|_2^2 + 2\|\mathbf{q}_2\|_2^2. && (14)\end{aligned}$$

Then

$$2nJ_{21} \leq C_{11}\|\mathbf{q}_1\|_2^2, \tag{15}$$

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

$$\begin{aligned} 2nJ_{22} &= 2\sum_{j=1}^{n}(1 - (1 - \beta\eta_j)^t)^2(v_j^T\epsilon)^2 \leq 2\sum_{j=1}^{n}\frac{4(\beta t\eta_j)^2}{(1 + \beta t\eta_j)^2}(v_j^T\epsilon)^2 && (16)\\ &= 8\epsilon^T(\tilde{\mathbf{K}} + (\beta t)^{-1}\mathbf{I})^{-1}\tilde{\mathbf{K}}^2(\tilde{\mathbf{K}} + (\beta t)^{-1}\mathbf{I})^{-1}\epsilon = 8\|\mathbf{q}_2\|_2^2, && (17)\end{aligned}$$

where $\eta_1 \geq ... \geq \eta_n > 0$ and v_j, $j = 1, ..., n$ be the eigenvalues and corresponding eigenvectors of $\tilde{\mathbf{K}}$, respectively. In the last inequality of (16), we note $(\beta t)^{-1} = n\lambda_n$.

Plugging (15) and (17) into (12), we have

$$J_2 \leq \frac{2C_{12}}{n}(\|\mathbf{q}_1\|_2^2 + \|\mathbf{q}_2\|_2^2), \tag{18}$$

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

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

and

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

where $\tilde{f}_n$ is as in Lemma 2 (statement of solution to optimization problem), and $\hat{f}_{0,n}$ is the solution to the optimization problem

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

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

$$\begin{aligned}
\lambda_n &\leq C_{13}(\lambda_n(m_\varepsilon+1)^{m_\varepsilon}\sigma_n^{2m_\varepsilon})^{\frac{m_f}{m_0+m_\varepsilon}} && (19)\\
&\Leftrightarrow n^{-\frac{2(m_0+m_\varepsilon)}{2m_f+d}} n^{\frac{4m_\varepsilon(2m_0+2m_\varepsilon)D-2m_\varepsilon(2m_0+2m_\varepsilon-D)d}{(2m_f+d)(4m_\varepsilon D-(2m_0+2(1-\delta_0)m_\varepsilon-D)d)}} \leq C_{14} n^{-\frac{2m_f}{2m_f+d}}(m_\varepsilon+1)^{\frac{m_\varepsilon m_f}{m_0+m_\varepsilon}} && (20)\\
&\Leftrightarrow m_\varepsilon^2\delta_0 d > m_\varepsilon(2m_f D + (m_0-m_f)(1-\delta_0)d) && (21)\\
&\Leftrightarrow m_\varepsilon > \frac{2m_f D + m_0 d}{\delta_0 d}, && (22)
\end{aligned}$$

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

$$\begin{aligned}
\lambda_n^{\frac{m_f}{m_0}} &\leq C_{15}(\lambda_n(m_\varepsilon+1)^{m_\varepsilon}\sigma_n^{2m_\varepsilon})^{\frac{m_f}{m_0+m_\varepsilon}} && (23)\\
&\Leftrightarrow n^{-\frac{2(m_0+m_\varepsilon)}{2m_f+d}} n^{\frac{4m_\varepsilon(2m_0+2m_\varepsilon)D-2m_\varepsilon(2m_0+2m_\varepsilon-D)d}{(2m_f+d)(4m_\varepsilon D-(2m_0+2(1-\delta_0)m_\varepsilon-D)d)}} \leq C_{16} n^{-\frac{2m_0}{2m_f+d}}(m_\varepsilon+1)^{\frac{m_\varepsilon m_0}{m_0+m_\varepsilon}} && (24)\\
&\Leftrightarrow m_\varepsilon^2\delta_0 d > 2m_0 m_\varepsilon D && (25)\\
&\Leftrightarrow m_\varepsilon > \frac{2m_0 D + m_0 d}{\delta_0 d}, && (26)
\end{aligned}$$

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}, with T as in Lemma 2. Suppose $D > 1$. Then calculation gives

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

for some positive constants C_{18}, where M_1 is as in Lemma 2, and

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

with the inequality because of (22) if $m_0 \leq m_f$ or (26) if $m_0 > m_f$. Therefore, taking $\delta_0 = d^{-1}(2m_f + d)a$ and $m_e = (\delta_0 d)^{-1}(2D\max(m_0, m_f) + m_0 d) + 1$, we have

$$\frac{1}{n}\|q_1\|_2^2 = \|f_n - f\|_2^2 = O_P\left(n^{-\frac{2m_f}{2m_f+d} + a}\right),$$

$$\frac{1}{n}\|q_2\|_2^2 = \|f_0, n - f_0\|_2^2 = O_P\left(n^{-\frac{2m_f}{2m_f+d} + a}\right).$$

Then by (18) and (26), we obtain

$$J_2 = O_P\left(n^{-\frac{2m_f}{2m_f+d} + a}\right),$$

which corresponds to the first statement of the theorem.

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

$$M_1 \leq C_{18} n^{-\frac{m_f}{2m_f+d} \cdot \frac{d}{2(2m_f+d)(\log n)}}(m_e + m_0)m^{r+\frac{1}{2}} \leq C_{19} n^{-\frac{m_f}{2m_f+d}}(m_e + m_0)m^{r+\frac{1}{2}},$$

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

$$\frac{1}{n}\|q_1\|_2^2 = \|f_n - f\|_2^2 = O_P\left(n^{-\frac{2m_f}{2m_f+d}(\log n)^{2m_f+1}}\right),$$

$$\frac{1}{n}\|q_2\|_2^2 = \|f_0, n - f_0\|_2^2 = O_P\left(n^{-\frac{2m_f}{2m_f+d}(\log n)^{2m_f+1}}\right).$$

Thus, by (18) and (28), we have

$$J_2 = O_P\left(n^{-\frac{2m_f}{2m_f+d}(\log n)^{2m_f+1}}\right),$$

which corresponds to the second statement of the Theorem.

It remains to bound $\|g_t - f^*\|_{L_2(\mathcal{P}_X)}$. Note that

$$\|g_t - f^*\|_{L_2(\mathcal{P}_X)} \leq \|g_t - f_t^*\|_{L_2(\mathcal{P}_X)} + \|f_n - f^*\|_{L_2(\mathcal{P}_X)} \leq \|g_t - f_n^*\|_{L_2(\mathcal{P}_X)} + T^{1/2},$$

and that

$$\|g_t - f_n^*\|_n \leq \|g_t - f^*\|_n + \|f_n - f^*\|_n \leq \|g_t - f^*\|_n + O_P\left(\left(T + n^{-1/2}T^{1/2}\right)^{1/2}\right)$$

$$\leq O_P\left(n^{-\frac{2m_f}{2m_f+d}}(\log n)^{2m_f+1}\right).$$

Therefore, it suffices to bound the difference between $\|g_t - f_n^*\|_{L_2(\mathcal{P}_X)}$ and $\|g_t - f_n^*\|_n$. By Lemma 2, we have

$$\|g_t\|_{\mathcal{N}_{\sigma_n}(\Omega)}^2 \leq \sigma_n^{-2m_0}\|g_t\|_{\mathcal{H}_{K_\sigma}(\Omega)}^2 \leq C_{20}\sigma_n^{-2m_0}\|g\|_{\mathcal{H}_{K_\sigma}(\Omega)}^2 = O_P\left(n^{\nu_1(\log n)^{2m_f+1}}\right),$$

for some positive constants C_{20}, where

$$\nu_1 = \frac{2(m_0 + m_e - m_f)}{2m_f + d} + 2(m_e - m_0)\nu,$$

and

$$\nu = -\frac{2(2m_0 + 2m_e)D - (2m_0 + 2m_e - D)d}{(2m_f + d)(4m_eD - (2m_0 + 2(1-\delta_0)m_e - D)d)}.$$

Consider function class $\mathcal{G} = \{h : h = (g_t - f_n^*)/(C_{21}\nu^{1/2}(\log n)^{m_f+1/2})\}$, 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 4 leads to

$$\|h_1\|_{L_\infty(\Omega)} \leq C_{22}\|h_1\|_{L_2(\mathcal{P}_X)}^{1-\frac{D}{2(m_0+m_e)}}\|h_1\|_{\mathcal{N}_\sigma(\Omega)}^{\frac{D}{2(m_0+m_e)}},$$

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

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

where $R_1 = \sup_{h_1\in\mathcal{G}}\|h_1\|_{L_2(\mathcal{P}_X)} \leq \sup_{h_1\in\mathcal{G}}\|h_1\|_{L_\infty(\Omega)} \leq \sup_{h_1\in\mathcal{G}}\|h_1\|_{\mathcal{N}_\sigma(\Omega)} < 1$, because of the reproducing property. Let $m = m_0+m_e$. Taking $c = C_{22}R_1^{1-\frac{D}{2m}} < 1$, and $\delta_n = C_{23}(\sigma_n^{-d}n^{-1}c_2m^{2m_D/(2m_D-D)})^{2m_D/(4m}$ for some positive constants C_{23} in Lemma 23, it can be checked that

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

for some positive constants C_{24}, which implies the conditions of Lemma 3 are fulfilled. Applying Lemma 3 to the case $\|g_t - f_n^*\|_{L_2(\mathcal{P}_X)}^2 \geq \delta_n^2 n\nu_1$, together with (29), we have

$$R_1 = O_P\left(\max\{n^{-\frac{m_f}{2m_f+d}-\nu_1/2}(\log n)^{m_f+1/2}, 8n\}\right).$$

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

$$R_1 \leq C_{26}(\sigma_n^{-d} n^{-1} c_2 m^{2m_D/(2m_D-D)})^{2m_D/(4m)},$$

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

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

where

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

If $\delta_n < n^{-\frac{m_f}{2m_f+d}-\nu_1/2}(\log n)^{m_f+1/2}$, then

$$R_1 = O_P\left(n^{-\frac{m_f}{2m_f+d}-\nu_1/2}(\log n)^{m_f+1/2}\right),$$

which implies

$$\|g_t - f^*\|_{L_2(\mathcal{P}_X)} = O_P\left(n^{-\frac{m_f}{2m_f+d}}(\log n)^{m_f+1/2}\right).$$

Here we note that the proof is still valid if we replace g_t with $\tilde{g}$. Therefore, in both cases we have

$$\|g_t - f^*\|_{L_2(\mathcal{P}_X)} = O_P\left(n^{-\frac{m_f}{2m_f+d}}(\log n)^{m_f+1/2}\right),$$

which, together with (17) and (5), finishes the proof.

□

We now provide the proof for the case with weight decay.

Proof. **With Weight Decay.** If $\alpha > 0$, we decompose the error by

$$\begin{aligned}\|f_t - f^*\|_{L_2(P_X)} \leq & \|f_t - g_t\|_{L_2(P_X)} + \|k(\cdot)^T(\alpha/\beta I + \tilde{K})^{-1}y - f^*\|_{L_2(P_X)} \\ & + \|\beta k(\cdot)^T((1-\alpha)I - \beta\tilde{K})^t(\alpha I + \beta\tilde{K})^{-1}y\|_{L_2(P_X)} \\ = & I_1 + I_2 + I_3. \end{aligned} \tag{33}$$

As in (5), there exists an N_0 (depending on n) such that when $N \geq N_0$,

$$I_1 = O_P\left(n^{-1/2}\right). \tag{34}$$

The second term is the error $\|\tilde{f}_n - f^*\|_{L_2(P_X)}$, where $\tilde{f}_n$ is as in Lemma 2. Lemma 2 gives us that

$$\|\tilde{f}_n - f^*\|_n = O_P\left(n^{-\frac{m_f}{2m_f+d}}\right).$$

It can be further shown that

$$I_2 = O_P\left(n^{-\frac{m_f}{2m_f+d}}\right), \tag{35}$$

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

It remains to bound I_3 in (33). By Cauchy-Schwarz inequality,

$$\begin{aligned}
&\|\beta k(\cdot)^T((1-\alpha)I - \beta\tilde{K})^t(\alpha I + \beta\tilde{K})^{-1}y\|_{L_2(P_X)} \\
&\leq \left\|\operatorname{tr}\left((\alpha/\beta I + \tilde{K})^{-1}yk(\cdot)^T\right)^2\right\|_{L_2(P_X)}^{1/2} \operatorname{tr}\left(((1-\alpha)I - \beta\tilde{K})^{2t}\right)^{1/2} \\
&\leq \|k(\cdot)^T(\alpha/\beta I + \tilde{K})^{-1}y\|_{L_2(P_X)} \operatorname{tr}\left(((1-\alpha)I - \beta\tilde{K})^{2t}\right)^{1/2} \\
&\leq \|k(\cdot)^T k(\cdot)\|_{L_2(P_X)}^{1/2}\|y\|_{L_2}/\alpha \\
&= O_P\left(n^{\frac{1+\frac{2(m_0+m_s)}{2m_f+d}}{2}}\sigma_n^{2m_e}(1-\alpha)^t\right).
\end{aligned} \tag{36}$$

Thus, there exists $t_0 > 0$ such that as long as $t > t_0$, I_2 dominates I_3. Combining (34), (35), and (36), we finish the proof.

□

Thus completes the proof on polynomial smoothing. We now prove a similar theorem for Gaussian smoothing.

Theorem 3 (Gaussian smoothing.)**.** *Suppose Assumptions 1, 2, 4 (C3), 5, and 6 are satisfied. Let $f_t(x)$ be $w_t^T k(x)$, with w_t^T the parameter obtained at the t-th iteration and $k(x) = (K_S(x - x_1), ..., K_S(x - x_n))^T$, let $\beta = n^{-1}C_1$ with the positive constant $C_1 \leq (2\sup_{x\in\mathbb{R}^D} K_S(x))^{-1}$, 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*

$$t \geq C_2 \left(\frac{m_f}{2m_f + d} + \frac{1}{2} \right) \frac{\log n}{\log(1-\alpha)}$$

for some positive constant C_2.

Then we have

$$\|f^* - \hat{f}_t\|_{L_2(P_X)}^2 = O_P \left(n^{-\frac{2m_f}{2m_f+d}} (\log n)^{D+1} \right), \tag{19}$$

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

Again, several lemmas are used in this proof. Their proofs can be found in the original paper and are omitted again for the sake of length.

Lemma 6. *Let* $k_\sigma(x - x')$ *be a Gaussian kernel defined by*

$$k_\sigma(x - x') = \exp\left(-\frac{\|x - x'\|_2^2}{4\sigma^2} \right), \tag{37}$$

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

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

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 *do not depend on* σ_n.

Lemma 7. *Let* f_n^* *be the solution to the optimization problem*

$$\min_{g \in \mathcal{H}_{\tilde{k}_S}(\Omega)} \|f^* - g\|_{L_2(P_X)}^2 + \lambda_n \|g\|_{\mathcal{H}_{\tilde{k}_S}(\Omega)}^2. \tag{38}$$

Then

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

and

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

for some positive constant C_3.

Lemma 8. *Let f_n^* be the solution to the optimization problem*

$$\min_{g \in \mathcal{H}_{\tilde{K}_S}(\Omega)} \|f^* - g\|^2_{L_2(P_X)} + \lambda_n \|g\|^2_{\mathcal{H}_{\tilde{K}_S}(\Omega)}. \tag{39}$$

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

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

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

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

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

$$M_1 = max\Bigg((T + n^{-1/2} T^{1/2})^{1/2},\ \sigma_n^{-\frac{d}{2} - \frac{pD}{4}}\, p^{-\frac{D+1}{2}}\, n^{-\frac{1}{2}}\, \lambda_n^{-\frac{p}{4}},$$

$$\lambda_n^{-\frac{p}{4-p}} \Big(\sigma_n^{-\frac{d}{2} - \frac{pD}{4}} p^{-\frac{D+1}{2}} n^{-\frac{1}{2}} (T + n^{-\frac{1}{2}} T^{\frac{1}{2}})^{\frac{1}{2} - \frac{p}{4}} \Big)^{\frac{2}{4-p}},$$

$$\Big(\sigma_n^{-\frac{d}{2} - \frac{pD}{4}} p^{-\frac{D+1}{2}} n^{-\frac{1}{2}} \big((\lambda_n^{-1} T)^{\frac{p}{2}} (T + n^{-\frac{1}{2}} T^{\frac{1}{2}})^{1 - \frac{p}{2}} \big) \Big)^{\frac{1}{2}},$$

$$(\sigma_n^{-\frac{d}{2} - \frac{pD}{4}} p^{-\frac{D+1}{2}} n^{-\frac{1}{2}})^{\frac{2}{2+p}} ((\lambda_n^{-1} T)^{\frac{p}{2+p}}) \Bigg).$$

$$M_2 = max\Bigg((\lambda_n^{-1} (T + n^{-\frac{1}{2}} T^{\frac{1}{2}}))^{\frac{1}{2}},\ \sigma_n^{-\frac{d}{2} - \frac{pD}{4}} p^{-\frac{D+1}{2}} n^{-\frac{1}{2}} \lambda_n^{-\frac{2+p}{4}},$$

$$(\lambda_n^{-1} \sigma_n^{-\frac{d}{2} - \frac{pD}{4}} p^{-\frac{D+1}{2}} n^{-\frac{1}{2}} (T + n^{-\frac{1}{2}} T^{\frac{1}{2}})^{\frac{1}{2} - \frac{p}{4}})^{\frac{2}{4-p}},$$

$$(\lambda_n^{-1} \sigma_n^{-\frac{d}{2} - \frac{pD}{4}} p^{-\frac{D+1}{2}} n^{-\frac{1}{2}} (\lambda_n^{-1} T)^{\frac{p}{2}} (T + n^{-\frac{1}{2}} T^{\frac{1}{2}})^{1 - \frac{p}{2}})^{\frac{1}{2}},$$

$$\lambda_n^{-\frac{1}{2}} (\sigma_n^{-\frac{d}{2} - \frac{pD}{4}} p^{-\frac{D+1}{2}} n^{-\frac{1}{2}})^{\frac{2}{2+p}} ((\lambda_n^{-1} T)^{\frac{p}{2+p}}) \Bigg).$$

Then we have

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

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

$$\min_{f \in \mathcal{H}_{\tilde{K}_S}(\Omega)} \|f^* - f\|_n^2 + \lambda_n \|f\|_{\mathcal{H}_{\tilde{K}_S}(\Omega)}^2, \tag{41}$$

then

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

Lemma 9 (Interpolation inequality for Gaussian RKHS)**.** *Let $g \in \mathcal{H}_\sigma(\mathbb{R}^D)$. For any $1 > r > 0$, we have*

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

where C_4 is a constant not depending on r, σ, or g.

We now proceed to the proof of the theorem, with and without weight decay.

Proof. **Without Weight Decay** We first decompose the error as

$$\|f_t - f^*\|_{L_2(P_X)} \leq \|f_t - g_t\|_{L_2(P_X)} + \|g_t - f^*\|_{L_2(P_X)}, \tag{42}$$

where g_t is as in the gradient update analysis (4.2).

By Lemma 5, the first term $\|f_t - g_t\|_{L_2(P_X)}$ in (42) can be bounded by

$$\|f_t - g_t\|_{L_2(P_X)} \leq C_5 \|f_t - g_t\|_{L_\infty(\Omega)} = O_P\left(\frac{n^2\sqrt{\log(N)/N}}{\eta(\tilde{K})^2}\right),$$

for some positive constant C_5, as long as

$$\frac{1}{2}\eta(\tilde{K}) \geq \nu\sqrt{\frac{\log N}{N}}. \tag{43}$$

Choose

$$N_0 = \frac{4n^2}{\eta(\tilde{K})^2}. \tag{44}$$

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

$$\|f_t - g_t\|_{L_2(P_X)} = O_P\big(n^{-1/2}\big). \tag{45}$$

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

$$J_2 = \|g_t - f^*\|_n^2 = \frac{1}{n}\|g_t(X) - f^*(X)\|_2^2. \tag{46}$$

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

$$\tilde{g} = \arg\min_{f \in \mathcal{H}_{K_S}(\Omega)} \|f - y\|_n^2 + \lambda_n \|f\|_{\mathcal{H}_{K_S}(\Omega)}^2.$$

By the Representer Theorem,

$$\tilde{g}(x) = \tilde{k}(x)^T(\tilde{K} + n\lambda_n I)^{-1} y$$

for all $x \in \Omega$. Then it can be seen that

$$\tilde{g}(X) - f^*(X) = n\lambda_n(\tilde{K} + n\lambda_n I)^{-1} f^*(X) + \tilde{K}(\tilde{K} + n\lambda_n I)^{-1}\varepsilon = q_1 + q_2.$$

Following the arguments in the previous theorem (proof without weight decay), the term J_2 can be bounded by

$$J_2 \leq \frac{2}{n}\big(2C_6\|q_1\|_2^2 + 8\|q_2\|_2^2\big), \tag{47}$$

for some positive constants C_6, and

$$\begin{aligned} \frac{1}{n}\|q_1\|_2^2 &= \|\tilde{f}_n - f^*\|_n^2, \\ \frac{1}{n}\|q_2\|_2^2 &= \|f_{0,n} - f_0\|_n^2, \end{aligned} \tag{48}$$

where $f_0(x) = 0$ for all $x \in \Omega$, $\tilde{f}_n$ is as in (41), and $f_{0,n}$ is the solution to the optimization problem

$$\min_{g \in \mathcal{H}_{K_S}(\Omega)} \|e - g\|_n^2 + \lambda_n \|g\|_{\mathcal{H}_{K_S}(\Omega)}^2.$$

By setting $\beta t \simeq n\lambda_n$ (which implies $\lambda_n \simeq n^{-\frac{2m_0}{2m_f+d}}$, $\sigma_n \simeq n^{-\frac{1}{2m_f+d}}$), Lemma 7 implies that $T \simeq n^{-\frac{2m_f}{2m_f+d}}$, which, together with Lemma 8, implies

$$\frac{1}{n}\|q_1\|_2^2 = \|\tilde{f}_n - f^*\|_n^2 = O_P\left(n^{-\frac{2m_f}{2m_f+d}}(\log n)^{D+1}\right),$$

and

$$\frac{1}{n}\|q_2\|_2^2 = \|f_{0,n} - f_0\|_n^2 = O_P\left(n^{-\frac{2m_f}{2m_f+d}}(\log n)^{D+1}\right).$$

By (48) and (47), we obtain

$$J_2 = O_P\big(n^{-\frac{2m_f}{2m_f+d}}(\log n)^{D+1}\big). \tag{49}$$

Next, we consider bounding $\|g_t - f^*\|_{L_2(P_X)}$. Similar to the proof of the first theorem (without weight decay), it suffices to consider bounding the difference between $\|g_t - f_n^*\|_{L_2(P_X)}$ and $\|g_t - f_n^*\|_n$. Lemma 6 implies that

$$\|\tilde{g}\|_{\lambda_n\sigma_n/\sqrt{2}(\Omega)} \leq C_7\sigma_n^{-D}\|\tilde{g}\|^2_{\mathcal{H}_{K_S}(\Omega)} = O_P\left(n^{-\frac{2m_0+2m_f}{2m_f+d}}(\log n)^{D+1}\right), \tag{50}$$

for some positive constants C_7.

Consider the function class

$$\mathcal{G} = \left\{h : h = (g_t - f_n^*)/(2C_8 n^{\frac{m_0+D/2}{2m_f+d}}(\log n)^{\frac{D+1}{2}})\right\}$$

where the constant C_8 is taken such that $\|h_1\|_{\lambda_n\sigma_n/\sqrt{2}(\Omega)} < 1$ for all $h_1 \in \mathcal{G}$. Taking $r = (\log n)^{-1}$ in Lemma 9, together with the extension theorem, leads to

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

□

Proof. **With Weight Decay** The proof is the same as the proof of Theorem 2 with weight decay, with the convergence rate for I_2 obtained from the proof (Theorem 3 without weight decay) above.

□

4.1 Additional Calculations for Proofs

4.1.1 Gradient Update Analysis

Let $\mathbf{X} = (x_1, ..., x_n)$, $\alpha > 0$ if there is weight decay, and $\alpha = 0$ if there is no weight decay. By the gradient update rule, we have

$$\begin{aligned} f_t(\mathbf{X}) &= \mathbf{K}w_t = \sqrt{\mathbf{K}}\theta_t \\ &= \sqrt{\mathbf{K}}\theta_t - \beta\sqrt{\mathbf{K}}(\theta_t - \sqrt{\mathbf{K}}y) - \alpha\sqrt{\mathbf{K}}\theta_t \\ &= ((1-\alpha)\mathbf{I} - \beta\mathbf{K})f_t(\mathbf{X}) + \beta\mathbf{K}y, \end{aligned}$$

which implies

$$\begin{aligned} f_{t+1}(\mathbf{X}) - \beta(\alpha\mathbf{I} + \beta\mathbf{K})^{-1}\mathbf{K}y &= ((1-\alpha)\mathbf{I} - \beta\mathbf{K})(f_t(\mathbf{X}) - \beta(\alpha\mathbf{I} + \beta\mathbf{K})^{-1}\mathbf{K}y) \\ &= ... = -((1-\alpha)\mathbf{I} - \beta\mathbf{K})^{t+1}\beta(\alpha\mathbf{I} + \beta\mathbf{K})^{-1}\mathbf{K}y, \end{aligned}$$

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

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

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

$$f_{t+1}(\mathbf{X}) = (\mathbf{I} - (\mathbf{I} - \beta\mathbf{K})^{t+1})y.$$

The estimator after t-th iteration can be obtained by

$$f_t(x) = w_t^T k(x) = k(x)^T \mathbf{K}^{-1} f_t(\mathbf{X}).$$

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 ε_{k_1} and ε_{k_2}, we define the expected smoothing kernel $\tilde{K}_S$ as

$$\tilde{K}_S(x, x') = \int_{\mathbb{R}^D} \int_{\mathbb{R}^D} K(x + \epsilon - (x' + \epsilon')) p_\varepsilon(e) p_\varepsilon(e')\, d\epsilon\, d\epsilon'. \tag{1}$$

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 $\tilde{K}_S$.

Let g_t be the function obtained at the t-th iteration by the gradient update rule with respect to the kernel function $\tilde{K}_S$. We have

$$g_t(X) = \tilde{K}(\alpha/\beta I + \tilde{K})^{-1}y - ((1-\alpha)I - \beta\tilde{K})\beta(\alpha I + \beta\tilde{K})^{-1}Ky, \tag{2}$$

if there is weight decay, and

$$g_t(X) = \big(I - (I - \beta\tilde{K})^t\big)y, \tag{3}$$

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

$$g_t(x) = \tilde{k}(x)^T \mathbf{K}^{-1} g_t(X). \tag{4}$$

Thus, the empirical error $\|f_t(X) - f^*(X)\|_2$ can be decomposed by

$$\|f_t(X) - f^*(X)\|_2 \le \|f_t(X) - g_t(X)\|_2 + \|g_t(X) - f^*(X)\|_2. \tag{5}$$