Learning Gradients via an Early Stopping Gradient Descent Method*

Xin Guo[a]

[a]*Department of Mathematics, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong, China.*

Abstract

We propose an early stopping algorithm for learning gradients. The motivation is to choose "useful" or "relevant" variables by a ranking method according to norms of partial derivatives in some function spaces. In the algorithm, we used an early stopping technique, instead of the classical Tikhonov regularization, to avoid over-fitting. The advantage includes that we need no longer consider the choice of a regularization parameter.

After stating dimension-dependent learning rates valid for any dimension of the input space, we present a novel error bound when the dimension is large. Our novelty is the independence of power index of the learning rates on the dimension of the input space.

Keywords:
gradient learning, early stopping, approximation error, reproducing kernel Hilbert spaces
2000 MSC: 62H30, 65C60

1. Introduction and Learning Algorithm

Variable and feature selection is a classical topic in statistics with the aim also of dimension reduction. A vast literature in learning theory addresses this issue. Recently, Mukherjee and Zhou [10] proposed a new method for variable selection with an idea of comparing norms of partial derivatives of an involved regression function in a regression setting. The learning algorithm was motivated by some applications from gene sequence analysis [7]. Following the work, Dong and Zhou

*The work described in this paper is supported partially by the Research Grants Council of Hong Kong [Project No. CityU 104007].

Email address: xinguo2@student.cityu.edu.hk (Xin Guo)

[5] used a gradient descent method to reduce computational complexity. Mukherjee and Wu [8] studied a general class of loss functions and constructed the corresponding efficient algorithms for classification problems. Mukherjee et. al. [9] studied the gradient learning problem on manifolds to capture the manifold property of the data spaces. The learning rates achieved in these results are low when the input space (or manifold) is of very high dimension. For example, in [5], the rate for any fixed confidence is $O(m^{-\theta})$ with $\theta \leq \frac{1}{6n+32}$, where m is the sample size and n is the dimension of the input space. Note that n is often very large for learning problems with dimension reduction or variable selection. The purpose of this paper is to study an early stopping algorithm for gradient learning. Our main novelty is that learning rates $O(m^{-\theta})$ achieved by our algorithm have power index θ independent of the input space dimension n when n is large. Such a dimension-independent learning rate has never appeared in the literature of gradient learning.

We set our input space X to be a compact subset of $\mathbb{R}^n$, and Y to be $\mathbb{R}$. Let $Z = X \times Y$, and ρ be a Borel probability measure on Z. We write ρ_X as the marginal distribution of ρ on X, and $\rho(y|x)$ the conditional distribution at $x = (x^1, \cdots, x^n) \in X$. Suppose we have a least square regression function $f_\rho(x) := \int_Y y \, \mathrm{d}\rho(y|x)$ which has almost everywhere the gradient

$$\nabla f_\rho(x) = \left(\frac{\partial f_\rho(x)}{\partial x^1}, \cdots, \frac{\partial f_\rho(x)}{\partial x^n} \right)^T \in \left(L^2_{\rho_X} \right)^n .$$

Our learning algorithm is a kernel method. The reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$ corresponding to a Mercer kernel (see Aronszajn [2]) K is defined as a completion of the linear span of the function set $\{K_x : K_x(\cdot) := K(x, \cdot)\}$ with respect to the inner product $\langle K_x, K_u \rangle_{\mathcal{H}_K} := K(x, u)$. We denote $\mathcal{H}_K^n := \{\vec{f} = (f_1, \cdots, f_n)^T : f_i \in \mathcal{H}_K, i = 1, \cdots, n\}$, then it is another Hilbert space with norm $\|\vec{f}\|_{\mathcal{H}_K^n} := \left(\sum_{i=1}^n \|f_i\|^2_{\mathcal{H}_K} \right)^{1/2}$, where $\| \cdot \|_{\mathcal{H}_K}$ is the norm on $\mathcal{H}_K$.

The risk functional for learning the gradient came from the Taylor expansion (see [10]): $f_\rho(u) \approx f_\rho(x) + \nabla f_\rho(x)^T \cdot (u - x)$ when $u \approx x$. So, to approximate ∇f_ρ by a vector valued function $\vec{f} = (f_1, \cdots, f_n)^T \in \mathcal{H}_K^n$, one method is to minimize the risk

$$\mathcal{E}(\vec{f}) = \int_X \int_X w^{(s)}(x, u)(f_\rho(x) - f_\rho(u) + \vec{f}(x)^T \cdot (u - x))^2 \, \mathrm{d}\rho_X(x) \, \mathrm{d}\rho_X(u) \quad (1.1)$$

in the space $\mathcal{H}_K^n$, where $w^{(s)}(x, u) > 0$ is the weight function to restrict $x \approx u$. Same as in [10], we require $w^{(s)}(x, u) \to 0$ as $\frac{x-u}{s} \to \infty$ to reduce the sample

error. In the following discussion we will use a special weight only

$$w(x,u) = w^{(s)}(x,u) = \frac{1}{s^{n+2}} \mathrm{e}^{-\frac{|x-u|^2}{2s^2}}.$$

For the sample set $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ drawn i.i.d. from ρ, denoting $w_{ij}^{(s)} = w^{(s)}(x_i, x_j)$, we have the empirical risk functional

$$\begin{aligned} \mathcal{E}^{\mathbf{z}}(\vec{f}) &= \frac{1}{m^2} \sum_{i,j=1}^m w_{ij}^{(s)} (y_i - y_j + \vec{f}(x_i)^T (x_j - x_i))^2 \\ &= \left\langle \vec{f}, L_{K,s}^{\mathbf{z}} \vec{f} \right\rangle_{\mathcal{H}_K^n} - 2 \left\langle \vec{f}_{\rho,s}^{\mathbf{z}}, \vec{f} \right\rangle_{\mathcal{H}_K^n} + C_0^{\mathbf{z}}, \end{aligned}$$

where $L_{K,s}^{\mathbf{z}} : \mathcal{H}_K^n \to \mathcal{H}_K^n$ is defined as

$$L_{K,s}^{\mathbf{z}} \vec{f} = \frac{1}{m^2} \sum_{i,j=1}^m w_{ij}^{(s)} (x_i - x_j)(x_i - x_j)^T \vec{f}(x_i) K_{x_i},$$

and

$$\begin{aligned} \vec{f}_{\rho,s}^{\mathbf{z}} &:= \frac{1}{m^2} \sum_{i,j=1}^m w_{ij}^{(s)} (y_i - y_j)(x_i - x_j) K_{x_i}, \\ C_0^{\mathbf{z}} &:= \frac{1}{m^2} \sum_{i,j=1}^m w_{ij}^{(s)} (y_i - y_j)^2. \end{aligned}$$

Now our learning algorithm can be expressed as

$$\vec{f}_{k+1}^{\mathbf{z}} = \vec{f}_k^{\mathbf{z}} - \gamma_k L_{K,s}^{\mathbf{z}} \vec{f}_k^{\mathbf{z}} + \gamma_k \vec{f}_{\rho,s}^{\mathbf{z}}, \quad k = 1, 2, \cdots, k^*, \tag{1.2}$$

where γ_k is the step size, having absorbed the constant 2. We set $\vec{f}_1^{\mathbf{z}} = 0$. The algorithm is called early stopping because the iteration procedure stops at step k^*. Instead of finding computational criteria for determining k^*, we shall conduct theoretical study on how a choice of type $k^* = m^a$ with $a > 0$ yields learning rates for gradient learning.

2. Main Results

We require some regularity of X, namely the cone property (see [1]), defined as

Definition 1. *A set $\Omega \subset \mathbb{R}^n$ has the cone property with parameter $0 < \varphi < \pi/2$ and $0 < R < +\infty$ if there exits a function $\vec{\alpha} : \Omega \to S^{n-1}$, such that for each $x \in \Omega$, the cone*

$$C_x = C_x(R, \varphi) := \{u \in \mathbb{R}^n : (u - x)^T \cdot \vec{\alpha}(x) > |u - x| \cos \varphi, |u - x| < R\}$$

is contained in Ω. □

In the following, we suppose that X satisfies the cone property, which in fact could be guaranteed by the Lipschitz condition of the boundary of X (denoted by ∂X). That is, for each $x \in \partial X$, there exists a neighborhood $U_x \subset \mathbb{R}^n$ such that $\partial X \cap U_x$ is the graph of a Lipschitz continuous function (of order 1) with some change of coordinates if necessary. Considering the compactness of X, we can thus bound the Lipschitz constant away from infinity. See [1], page 66-67.

We here use the Mercer kernel K defined on $X \times X$, thus the RKHS $\mathcal{H}_K$ is contained in $L^2_{\rho_X} \cap C(X)$. We define $L_K : \left(L^2_{\rho_X}\right)^n \to \left(L^2_{\rho_X}\right)^n$ as

$$(L_K \vec{f})(u) := \int_X \vec{f}(x) K(x, u) \, \mathrm{d}\rho_X(x),$$

thus L_K becomes a positive operator on $\left(L^2_{\rho_X}\right)^n$.The range of L_K lies in $\mathcal{H}^n_K$ and the restriction of L_K onto $\mathcal{H}^n_K$ is also positive. Besides, $L_K^{1/2}\left(\left(L^2_{\rho_X}\right)^n\right) \subset \mathcal{H}^n_K$ and $\|\vec{f}\|_\rho = \|L_K^{1/2}\vec{f}\|_{\mathcal{H}_K}$ for $\vec{f} \in L_K^{1/2}\left(\left(L^2_{\rho_X}\right)^2\right)$, where $\|\cdot\|_\rho$ is the canonical norm defined on $\left(L^2_{\rho_X}\right)^n$ as

$$\|\vec{f}\|_\rho := \left(\sum_{k=1}^n \int_X f_k^2(x) \, \mathrm{d}\rho_X(x)\right)^{1/2}.$$

We suppose that $\nabla f_\rho \in L_K\left(\left(L^2_{\rho_X}\right)^n\right) \subset \mathcal{H}^n_K$, then

$$\|\nabla f_\rho\|_\infty := \operatorname*{ess\,sup}_{x \in X} \left(\sum_{i=1}^n \left(\frac{\partial f_\rho(x)}{\partial x^i}\right)^2\right)^{1/2}$$

exists and it is finite. Denote $\kappa := \sup_{x \in X} \sqrt{K(x, x)} < +\infty$.

Denote $J_p = \int_{\mathbb{R}^n} |x|^2 \mathrm{e}^{-\frac{|x^2|}{2}} \, \mathrm{d}x$ for $p \geq 0$. For learning gradients we assume throughout the paper that ρ_X has a C^1 density function p on X and we write $c_p = \|p\|_{C(X)}$. The case $n = 1$ is omitted because it is trivial for ranking based variable selection problems.

Theorem 1. *Let $n \geq 2$ and $0 \leq \tau < 1$. Assume that X satisfies the cone property with parameters (R, φ), and $p(x) \geq \beta(\mathrm{dist}(x, \partial X))^{\alpha}$ for some $\beta > 0$ and $0 \leq \alpha < 3/2$. Take the iteration step size $\gamma_t = \gamma_1 t^{-\tau}$ with $\gamma_1 = \frac{s^n}{\kappa^2(1+c_p J_2)}$. If $|y| \leq M$ almost surely and ∇f_ρ has the regularity that $\nabla f_\rho \in L_K\left(\left(L^2_{\rho_X}\right)^n\right)$, then by taking the weight parameter $s = s_0 m^{-1/(4n+11-2\alpha)}$ and the step $k^* = k^*(m) = \left\lceil \left((1-\tau)m^{(n+\frac{3}{2})/(4n+11-2\alpha)}\right)^{1/(1-\tau)} \right\rceil - 1$, for $m > (1-\tau)^{\frac{4\alpha-22-8n}{2n+3}}$, we have with confidence $1-\delta$ for $0 < \delta < 1$,*

$$\|\vec{f}^{\mathbf{z}}_{k^*+1} - \nabla f_\rho\|_\rho \leq C_1 m^{-\frac{\frac{3}{2}-\alpha}{11+4n-2\alpha}} \left(1 + \frac{\log m}{4(1-\tau)}\right) \log \frac{4}{\delta},$$

with $s_0 = \min\left\{1, \frac{R}{3\sqrt{n+4}}\right\}$, and C_1 is a constant independent of m or δ.

The learning rates in Theorem 1 depend on the dimension n of the input space. The corresponding power index $-\frac{\frac{3}{2}-\alpha}{11+4n-2\alpha}$ is very small when n is large. Similar rates were achieved in [5, 10]. Meanwhile, learning rates in classical results (e.g. [12, 13]) of least square regression learning by kernel methods do not have the shortcoming. The upper bounds achieved in [12, 13] of the rates are independent of the input space dimension. To achieve such a dimension-independent learning rate, we give the following theorem.

Theorem 2. *Let $n \geq 23$. Under the same conditions as in Theorem 1, take the step size $\gamma_t = \gamma_1 t^{-\tau}$ and $0 < \gamma_1 \leq (\sqrt{n}\kappa^2 c_p J_2)^{-1}$. By taking the weight parameter $s = s^{\mathbf{z}}_0 m^{\frac{-1}{11-2\alpha}+\frac{2}{n}}$ with*

$$\begin{aligned} s^{\mathbf{z}}_0 &= \min\left\{1, \frac{R}{3\sqrt{n+4}}, \frac{\varepsilon_{\mathbf{z}}}{\frac{2(n+2)}{\mathrm{e}} + \sqrt{2|\log(\varepsilon^n_{\mathbf{z}} c_p J_2)|}}\right\}, \\ \varepsilon_{\mathbf{z}} &= \min\{|x_i - x_j| : 1 \leq i < j \leq m\}, \end{aligned}$$

and the step $k^* = k^*(m) = \left\lceil \left((1-\tau)m^{\frac{3/2}{11-2\alpha}}\right)^{1/(1-\tau)} \right\rceil - 1$, for $m > (1-\tau)^{\frac{2\alpha-11}{3/2}}$, we have with confidence $1-\delta$ for $0 < \delta < 1$, that

$$\|\vec{f}^{\mathbf{z}}_{k^*+1} - \nabla f_\rho\|_\rho \leq C_2 m^{-\left(\frac{3}{2}-\alpha\right)\left(\frac{1}{11-2\alpha}-\frac{2}{n}\right)} \left(1 + 2\sqrt{\frac{\log m}{1-\tau}}\right)^2 \delta^{-\frac{3}{2n}} \left(\log \frac{6\mathrm{e}}{\delta}\right)^{\frac{5}{2}},$$

where C_2 is a constant independent of m or δ.

Remark. The power index $\left(\frac{3}{2}-\alpha\right)\left(\frac{1}{11-2\alpha}-\frac{2}{n}\right) \geq \left(\frac{3}{2}-\alpha\right)\left(\frac{1}{11-2\alpha}-\frac{2}{23}\right)$ is independent of the dimension n. Though δ appears in a form of polynomial, the power $\frac{3}{2n}$ is very small when n is large.

3. Structure of Integral Operators

The gradient descent algorithm, although simple and economic in computation, does not always provide satisfactory convergence rates. In some cases we cannot guarantee the convergence at all since there may exist some directions to which the risk function (or functional) could be very flat, having the principle curvature hard to be bounded away from zero. This problem could be solved by adding the Tikhonov regularization term as done in [15] and [5], with the shortcoming that the regularization parameter may sometimes be difficult to fix, as well as that bias may be introduced. Another way, called the early stopping method, as in [14], is to exploiting more properties of the gradient, and prove that during the whole process of iterations, one never goes through those directions of low curvature. We will use the early stopping method, the shortcoming of which, as will be shown below, is that we have to compose more prior assumptions, which might restrict its applicability.

Let us define a sample-free limit of algorithm (1.2). We rewrite (1.1) as a quadratic functional in $\mathcal{H}_K^n$:

$$\mathcal{E}(\vec{f}) = \left\langle \vec{f}, L_{K,s}\vec{f} \right\rangle_{\mathcal{H}_K^n} - 2\left\langle \vec{f}_{\rho,s}, \vec{f} \right\rangle_{\mathcal{H}_K^n} + C_0, \tag{3.1}$$

where $L_{K,s} : (L^2_{\rho_X})^n \to \mathcal{H}_K^n$ is defined as (see [10])

$$L_{K,s}\vec{f} = \int_X \int_X w(x,u)(u-x)(u-x)^T \vec{f}(x) K_x \,\mathrm{d}\rho_X(u)\,\mathrm{d}\rho_X(x),$$

and

$$\begin{aligned} \vec{f}_{\rho,s} &= \int_X \int_X w(x,u)(f_\rho(u) - f_\rho(x))(u-x) K_x \,\mathrm{d}\rho_X(u)\,\mathrm{d}\rho_X(x), \\ C_0 &= \int_X \int_X w(x,u)(f_\rho(x) - f_\rho(u))^2 \,\mathrm{d}\rho_X(u)\,\mathrm{d}\rho_X(x). \end{aligned}$$

So we can take gradient of $\mathcal{E}(\vec{f})$ in $\mathcal{H}_K^n$

$$\mathrm{Grad}\mathcal{E}(\vec{f}) = 2(L_{K,s}\vec{f} - \vec{f}_{\rho,s}),$$

and thus we get the so-called population iteration scheme for minimizing $\mathcal{E}(\vec{f})$ in $\mathcal{H}_K^n$:

$$\vec{f}_{k+1} = \vec{f}_k - \frac{1}{2}\gamma_k \mathrm{Grad}\mathcal{E}(\vec{f}_k) = \vec{f}_k - \gamma_k L_{K,s}\vec{f}_k + \gamma_k \vec{f}_{\rho,s}, \;\; k = 1, 2, \cdots.$$

We set $\vec{f}_1 = 0$.

From the reproducing property

$$f(x) = \langle f, K_x \rangle_{\mathcal{H}_K}, \ \ \forall f \in \mathcal{H}_K, \, x \in X,$$

we see that $|f(x)| \leq \|f\|_{\mathcal{H}_K} \|K_x\|_{\mathcal{H}_K} \leq \kappa \|f\|_{\mathcal{H}_K}$. Hence

$$\|\vec{f}\|_\rho \leq \kappa \|\vec{f}\|_{\mathcal{H}_K^n}. \tag{3.2}$$

We define $\Gamma_s : (L^2_{\rho_X})^n \to (L^2_{\rho_X})^n \ \ (\forall s > 0)$ as

$$(\Gamma_s \vec{f})(x) := \int_X w(x,u)(u-x)(u-x)^T \, \mathrm{d}\rho_X(u) \vec{f}(x).$$

Obviously, Γ_s is a positive operator. For any $\vec{f} \in (L^2_{\rho_X})^n$,

$$\begin{aligned} L_K \Gamma_s \vec{f} &= \int_X K_x \vec{f}(x) \, \mathrm{d}\rho_X(x) \int_X w(x,u)(u-x)(u-x)^T \, \mathrm{d}\rho_X(u) \\ &= \int_X \int_X w(x,u)(u-x)(u-x)^T \vec{f}(x) K_x \, \mathrm{d}\rho_X(x) \, \mathrm{d}\rho_X(u) \\ &= L_{K,s} \vec{f}, \end{aligned}$$

that is, $L_{K,s} = L_K \circ \Gamma_s$. Unfortunately, L_K and Γ_s do not commute, so generally we can not represent $L^r_{K,s}$ by $L^r_K \circ \Gamma^r_s$. But Γ_s is invertible when s is sufficiently small. For proving this, we need a lemma.

Lemma 1. *For $a \geq \frac{1}{2}$,*

$$\frac{1}{\Gamma(a)} \int_{2a}^{+\infty} \mathrm{e}^{-y} y^{a-1} \, \mathrm{d}y \leq \frac{1}{2},$$

where $\Gamma(a)$ is the Gamma function defined by $\Gamma(a) = \int_0^{+\infty} \mathrm{e}^{-y} y^{a-1} \, \mathrm{d}y$.

Proof. For $a \geq \frac{1}{2}$, denote

$$I = \int_{2a}^{+\infty} \mathrm{e}^{-y} y^{a-1} \, \mathrm{d}y = a^a \int_2^{+\infty} \mathrm{e}^{-at} t^{a-1} \, \mathrm{d}t.$$

Let $\mathrm{e}^{-1-u} = t\mathrm{e}^{-t}$, we have $u = t - \log t - 1$ and $\frac{\mathrm{d}t}{t} = \frac{\mathrm{d}u}{t-1}$. Note that $t \geq 2$, so

$$I = a^a \int_{1-\log 2}^{+\infty} \mathrm{e}^{-a} \mathrm{e}^{-au} \frac{1}{t-1} \, \mathrm{d}u \leq a^a \mathrm{e}^{-a} \int_{1-\log 2}^{+\infty} \mathrm{e}^{-au} \, \mathrm{d}u = a^{a-1} \mathrm{e}^{-a(2-\log 2)}.$$

By the Stirling's formula,

$$\Gamma(a) \geq \sqrt{2\pi} a^{a-\frac{1}{2}} \mathrm{e}^{-a},$$

because $a \geq \frac{1}{2}$,

$$\frac{1}{\Gamma(a)} \int_{2a}^{+\infty} \mathrm{e}^{-y} y^{a-1} \, \mathrm{d}y \leq \frac{\mathrm{e}^{-a(1-\log 2)}}{\sqrt{2\pi a}} \leq \frac{\sqrt{2}}{\sqrt{\pi \mathrm{e}}} < \frac{1}{2},$$

which completes the proof. □

Theorem 3. *If the density function of ρ_X satisfies the boundary condition: there exists $\beta > 0$ and $0 \leq \alpha < 3/2$ such that $p(x) \geq \beta(\mathrm{dist}(x, \partial X))^\alpha$, and if X has the cone property with parameters (R, φ) as was mentioned before, then Γ_s is invertible for $0 < s \leq \frac{R}{3\sqrt{n+4}}$, and $\|\Gamma_s^{-1}\| \leq \frac{1}{s^\alpha w}$, where*

$$w = w(\alpha, \beta, R, \varphi) := \beta \frac{\pi^{n/2} \Gamma(\frac{n+\alpha+2}{2})}{\Gamma(\frac{n-2}{2})} 2^{\frac{n+\alpha}{2}-3} (\varphi - \sin\varphi) \sin^\alpha \frac{\varphi}{2}. \quad (3.3)$$

Proof. For any vector $\xi \in \mathbb{R}^n$ and $x \in X$, we claim that for $0 < s \leq \frac{R}{3\sqrt{n+4}}$,

$$I_1 = \int_X w(u, x) \left((u-x)^T \xi\right)^2 p(u) \, \mathrm{d}u \geq s^\alpha w |\xi|^2.$$

Note that

$$I_1 \geq \int_{C_x} w(u, x) \left((u-x)^T \xi\right)^2 \beta \left(\mathrm{dist}(u, \partial C_x)\right)^\alpha \, \mathrm{d}u.$$

Without loss of generality, we set $x = 0, C_x = C_0(R, \varphi) = \{u \in \mathbb{R}^n : u^T e_1 > |u| \cos\varphi, |u| < R\}$, and $\xi = |\xi| \cos\psi e_1 + |\xi| \sin\psi e_2$, where $e_1 = (1, 0, \cdots, 0)^T, e_2 = (0, 1, 0, \cdots, 0)^T \in \mathbb{R}^n$. We use the standard polar coordinates for $u = (u_1, \cdots, u_n)^T \in \mathbb{R}^n$: $u_1 = t\cos\varphi_1, u_2 = t\sin\varphi_1\cos\varphi_2, \cdots, u_n = t\sin\varphi_1\sin\varphi_2\cdots\sin\varphi_{n-1}$. We write $a = u^T e_1, b = \sqrt{|u|^2 - a^2}$, so for any $u \in C_0$, $\mathrm{dist}(u, \partial C_0) = \min\{R - |u|, a\sin\varphi - b\cos\varphi\}$, thus when $R - |u| \geq a\sin\varphi - b\cos\varphi$, or sufficiently when $|u| \leq \frac{R}{2} \leq \frac{R}{1+\sin\varphi}$, $\mathrm{dist}(u, \partial X) \geq a\sin\varphi - b\cos\varphi$. We have

$$I_1 \quad \geq \quad \int_{C_0(\frac{R}{2}, \varphi)} \frac{1}{s^{n+2}} \mathrm{e}^{-\frac{|u|^2}{2s^2}} \left(u^T \xi\right)^2 \beta \left(a\sin\varphi - b\cos\varphi\right)^\alpha \, \mathrm{d}u,$$

which equals

$$\beta s^\alpha |\xi|^2 \int_0^{\frac{R}{2s}} t^{n+1+\alpha} \mathrm{e}^{-\frac{t^2}{2}} \, \mathrm{d}t \int_0^{2\pi} \mathrm{d}\varphi_{n-1} \int_0^\pi \sin\varphi_{n-2} \, \mathrm{d}\varphi_{n-2} \cdots \int_0^\pi (\sin\varphi_3)^{n-4} \, \mathrm{d}\varphi_3$$
$$\cdot \int_0^\varphi \mathrm{d}\varphi_1 \int_0^\pi (\cos\varphi_1 \cos\psi + \sin\varphi_1 \cos\varphi_2 \sin\psi)^2 \sin^\alpha(\varphi - \varphi_1) \, \mathrm{d}\varphi_2.$$

Hence

$$\begin{aligned} I_1 &\geq \beta s^\alpha |\xi|^2 \frac{2\sqrt{\pi}^{n-2}}{\Gamma(\frac{n-2}{2})} \int_0^{\frac{R}{2s}} t^{n+1+\alpha} \mathrm{e}^{-\frac{t^2}{2}} \,\mathrm{d}t \\ &\quad \cdot \int_0^{\varphi} \sin^\alpha(\varphi-\varphi_1)\,\mathrm{d}\varphi_1 \int_0^{\pi} (\cos^2\varphi_1 \cos^2\psi + \sin^2\varphi_1 \sin^2\psi \cos^2\varphi_2)\,\mathrm{d}\varphi_2, \end{aligned}$$

which implies

$$\begin{aligned} I_1 &\geq \beta s^\alpha |\xi|^2 \frac{2\sqrt{\pi}^{n-2}}{\Gamma(\frac{n-2}{2})} 2^{\frac{n+\alpha}{2}} \int_0^{\frac{R^2}{8s^2}} y^{\frac{n+\alpha}{2}} \mathrm{e}^{-y}\,\mathrm{d}y \\ &\quad \cdot \pi \int_0^{\varphi} \sin^\alpha(\varphi-\varphi_1)(\cos^2\varphi_1\cos^2\psi + \frac{1}{2}\sin^2\varphi_1\sin^2\psi)\,\mathrm{d}\varphi_1. \end{aligned}$$

When $0 \leq \varphi_1 \leq \frac{\varphi}{2} < \frac{\pi}{4}$, $\frac{3}{2}\sin^2\varphi_1 - 1 \leq \frac{3}{4} - 1 < 0$, so

$$\begin{aligned} \cos^2\varphi_1\cos^2\psi + \frac{1}{2}\sin^2\varphi_1\sin^2\psi &= \cos^2\varphi_1 + \sin^2\psi\left(\frac{3}{2}\sin^2\varphi_1 - 1\right) \\ &\geq \cos^2\varphi_1 + \frac{3}{2}\sin^2\varphi_1 - 1 = \frac{1}{2}\sin^2\varphi_1, \end{aligned}$$

and also, when $s \leq \frac{R}{3\sqrt{n+4}}$, we have $\frac{R^2}{8s^2} \geq 2\left(\frac{n+\alpha}{2}+1\right)$, so by Lemma 1,

$$\begin{aligned} I_1 &\geq \beta s^\alpha |\xi|^2 \frac{2^{\frac{n+\alpha}{2}}\pi^{n/2}}{\Gamma\left(\frac{n-2}{2}\right)} \Gamma\left(\frac{n+\alpha+2}{2}\right)\left(\sin^\alpha\frac{\varphi}{2}\right)\int_0^{\varphi/2} \frac{1}{2}\sin^2\varphi_1\,\mathrm{d}\varphi_1 \\ &= \beta s^\alpha |\xi|^2 \frac{\pi^{n/2}\Gamma\left(\frac{n+\alpha+2}{2}\right)}{\Gamma\left(\frac{n-2}{2}\right)} 2^{\frac{n+\alpha}{2}-3}(\varphi - \sin\varphi)\sin^\alpha\frac{\varphi}{2} = s^\alpha w|\xi|^2. \end{aligned}$$

This verifies our claim.

For any $\vec{f} \in \left(L^2_{\rho_X}\right)^n$,

$$\begin{aligned} \|\Gamma_s\vec{f}\|_\rho^2 &= \int_X \left|\int_X w(x,u)(u-x)(u-x)^T\,\mathrm{d}\rho_X(u)\vec{f}(x)\right|^2 \mathrm{d}\rho_X(x) \\ &\leq \int_X \left|\int_X w(x,u)|u-x|^2 p(u)\,\mathrm{d}u\right|^2 \left|\vec{f}(x)\right|^2 \mathrm{d}\rho_X(x) \\ &\leq J_2^2 c_p^2 \|\vec{f}\|_\rho^2, \end{aligned}$$

so Γ_s is bounded and

$$\|\Gamma_s\| \leq c_p J_2. \tag{3.4}$$

On the other hand, we have

$$\left\langle \Gamma_s \vec{f}, \vec{f} \right\rangle_\rho \geq s^\alpha w \int_X |\vec{f}(x)|^2 \, \mathrm{d}\rho_X(x) = s^\alpha w \|\vec{f}\|_\rho^2,$$

which implies the conclusion. □

Remark: We find from the proof that the lower bound with parameters (α, β), $p(x) \geq \beta(\mathrm{dist}(x, \partial X))^\alpha$ could be replaced by

$$p(x) \geq \beta(\mathrm{dist}(x, \partial C_y))^\alpha \tag{3.5}$$

for any $y \in X$ and $x \in C_y$. Condition (3.5) keeps the assumption away from the severe requirement of $p(x)$ when x is far away from the boundary ∂X.

4. Sample Error

The main results in this section are Lemma 2 and Lemma 3, which are for proving Theorem 1 and Theorem 2 respectively. In the following analysis we suppose that $|y| \leq M < \infty$ almost surely. M is also used as an upper bound of $\frac{1}{2}\|\nabla f_\rho\|_\infty$ for saving the notations.

Lemma 2. *Let $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ be drawn independently from (Z, ρ), and $0 < \gamma_1 \leq \left(\frac{\kappa^2(1+c_p J_2)}{s^n}\right)^{-1}$, for any $\delta \in (0, \frac{1}{2})$ and any $s > 0$, we have with confidence $1 - 2\delta$:*

$$\|\vec{f}_{k+1}^{\mathbf{z}} - \vec{f}_{k+1}\|_{\mathcal{H}_K^n} \leq \frac{C_3(k+1)^{2-2\tau}}{\sqrt{m}s(1-\tau)^2} \log\frac{2}{\delta}, \tag{4.1}$$

where

$$C_3 = \frac{34M}{\kappa\sqrt{\mathrm{e}}}\left(\frac{\sqrt{n}}{\mathrm{e}} + 1\right).$$

Lemma 3. *Let $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ be drawn independently from (Z, ρ). Set $0 < \gamma_1 \leq \left(\sqrt{n}\kappa^2 c_p J_2\right)^{-1}$. For any $\delta \in (0, 1/3)$, we have with confidence at least $1 - 3\delta$ the estimation*

$$\|\vec{f}_{k+1}^{\mathbf{z}} - \vec{f}_{k+1}\|_{\mathcal{H}_K^n} \leq \frac{91M\left(2\sqrt{\log m} + 1\right)(k+1)^{2-2\tau}}{\kappa\sqrt{n}C_4\delta^{1/n}(1-\tau)^2} m^{\frac{2\alpha-9}{2(11-2\alpha)}} \left(\log\frac{2\mathrm{e}}{\delta}\right)^{\frac{5}{2}}, \tag{4.2}$$

where we define the weight parameter $s = s_0^{\mathbf{z}} m^{\frac{-1}{11-2\alpha}+\frac{2}{n}}$ with

$$s_0^{\mathbf{z}} = \min\left\{1, \frac{R}{3\sqrt{n+4}}, \frac{\varepsilon_{\mathbf{z}}}{\frac{2(n+2)}{\mathrm{e}} + \sqrt{2|\log(\varepsilon_{\mathbf{z}} c_p J_2)|}}\right\},$$

and C_4 is a constant depending only on (X, ρ_X).

Lemma 2 and Lemma 3 will be proved later in this section.

A linear bounded operator L on a Hilbert space H is said to be a Hilbert-Schmidt operator if for an orthonormal basis $\{e_i\}_{i\in I}$ of H, one has $\|L\|_{\mathrm{HS}} := \left(\sum_{i\in I}\|Le_i\|_H^2\right)^{1/2} < +\infty$. It can be proved that the Hilbert Schmidt norm $\|\cdot\|_{\mathrm{HS}}$ is independent of the choice of the basis $\{e_i\}_{i\in I}$. Any finite rank operator is a Hilbert-Schmidt operator. For any self-adjoint Hilbert-Schmidt operator L, one has $\|L\|_{\mathrm{HS}} \geq \|L\|$.

In the Hilbert space $\mathcal{H}_K^n$, we define for any $x \in X$, $A_x : \vec{f} \mapsto \vec{f}(x)K_x$. Then

$$\|A_x\vec{f}\|_{\mathcal{H}_K^n}^2 \leq \sum_{i=1}^n \|f_i\|_{\mathcal{H}_K}^2\|K_x\|_{\mathcal{H}_K}^4 = K(x,x)^2\|\vec{f}\|_{\mathcal{H}_K^n}^2 \leq \kappa^4\|\vec{f}\|_{\mathcal{H}_K}^2.$$

So $\|A_x\| \leq \kappa^2$. Also, it is obvious that A_x is self-adjoint. Since the rank of A_x is no greater than n, it is a Hilbert Schmidt operator. Let $e_1, \cdots, e_q \in \mathcal{H}_K^n$ be an orthonormal set spanning the range of A_x. So one has $\|A_x\|_{\mathrm{HS}}^2 = \sum_{i=1}^q \|Ae_i\|_{\mathcal{H}_K^n}^2 \leq q\kappa^4 \leq n\kappa^4$, and thus $\|A_x\|_{\mathrm{HS}} \leq \sqrt{n}\,\kappa^2$ for any $x \in X$. We have the relations

$$L_{K,s} = \int_X\int_X w(x,u)(u-x)(u-x)^T A_x \,\mathrm{d}\rho_X(u)\,\mathrm{d}\rho_X(x) \tag{4.3}$$

$$L_{K,s}^{\mathbf{z}} = \frac{1}{m^2}\sum_{i,j=1}^m w_{i,j}^{(s)}(x_i - x_j)(x_i-x_j)^T A_{x_i}. \tag{4.4}$$

$L_{K,s}$ and $L_{K,s}^{\mathbf{z}}$ are both self-adjoint. Also, from (4.3) and (4.4) one can directly compute

$$\begin{aligned}\|L_{K,s}\|_{\mathrm{HS}} &\leq \int_X\int_X w(x,u)|u-x|^2\|A_x\|_{\mathrm{HS}}\,\mathrm{d}\rho_X(x)\,\mathrm{d}\rho_X(u)\\ &\leq \sqrt{n}\kappa^2 c_p J_2.\end{aligned} \tag{4.5}$$

Moreover, $\mathbb{E}L_{K,s}^{\mathbf{z}} = \frac{m-1}{m}L_{K,s}$, and similarly, $\mathbb{E}\vec{f}_{\rho,s}^{\mathbf{z}} = \frac{m-1}{m}\vec{f}_{\rho,s}$.

Preparing for proving Lemma 2, we cite the following lemma from [5] with a little refinement, which could be done as noticing that $\frac{1}{s^{n+2}}\mathrm{e}^{-\frac{v^2}{2s^2}}v^2 \leq \frac{2}{\mathrm{e}s^n}$ and $\frac{1}{s^{n+2}}\mathrm{e}^{-\frac{v^2}{2s^2}}v \leq \frac{1}{\sqrt{\mathrm{e}}s^{n+1}}$ for any $v \in \mathbb{R}$, during the proof in [5].

Lemma 4. *Let $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ be independently drawn from (Z,ρ), and $L_{K,s}^{\mathbf{z}}$, $L_{K,s}$, $\vec{f}_{\rho,s}^{\mathbf{z}}$, $\vec{f}_{\rho,s}$ be defined as before. For any $s > 0$ and any $\delta \in (0, 1/2)$, with confidence $1-2\delta$, the following inequalities hold,*

$$\begin{aligned}\|L_{K,s}^{\mathbf{z}} - L_{K,s}\|_{\mathrm{HS}} &\leq \frac{34\kappa^2\sqrt{n}\log\frac{2}{\delta}}{\mathrm{e}\sqrt{m}s^n}\\ \|\vec{f}_{\rho,s}^{\mathbf{z}} - \vec{f}_{\rho,s}\|_{\mathcal{H}_K^n} &\leq \frac{34M\kappa\log\frac{2}{\delta}}{\sqrt{\mathrm{e}m}s^{n+1}}.\end{aligned}$$

□

Lemma 5. *Let* $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ *be independently drawn from* (Z, ρ) *and* $s > 0$*, then*

$$\|\vec{f}^{\mathbf{z}}_{\rho,s}\|_{\mathcal{H}_K^n} \leq \frac{2\kappa M}{\sqrt{\mathrm{e}} s^{n+1}}, \tag{4.6}$$

and

$$\|L^{\mathbf{z}}_{K,s}\|_{\mathfrak{L}(\mathcal{H}_K^n)} \leq \frac{2\kappa^2}{\mathrm{e} s^n} \tag{4.7}$$

hold almost surely.

Proof. (4.6) follows directly from the definition. (4.7) holds because

$$\begin{aligned} \left| \left\langle L^{\mathbf{z}}_{K,s} \vec{f}, \vec{f} \right\rangle_{\mathcal{H}_K^n} \right| &= \frac{1}{m^2} \sum_{i,j=1}^n w_{ij}^{(s)} \left((x_i - x_j)^T \vec{f}(x_i) \right)^2 \\ &\leq \frac{m-1}{m} \frac{2}{\mathrm{e} s^n} \kappa^2 \|\vec{f}\|^2_{\mathcal{H}_K^n} \end{aligned}$$

for any $\vec{f} \in \mathcal{H}_K^n$. □

We see that if we set $0 < \gamma_1 \leq \left(\frac{\kappa^2(1+c_p J_2)}{s^n} \right)^{-1}$, then for $t = 1, 2, \cdots$, $\|\gamma_t L^{\mathbf{z}}_{K,s}\|_{\mathfrak{L}(\mathcal{H}_K^n)} \leq 1$ almost surely.

Lemma 6. *Let* $s > 0$*, for* $\vec{f}^{\mathbf{z}}_k$ *recurrently defined in (1.2) and* $0 < \gamma_1 \leq \left(\frac{\kappa^2(1+c_p J_2)}{s^n} \right)^{-1}$*, we have*

$$\|\vec{f}^{\mathbf{z}}_k\|_{\mathcal{H}_K^n} \leq \frac{2\kappa M \gamma_1 k^{1-\tau}}{\sqrt{\mathrm{e}} s^{n+1}(1-\tau)}, \quad k = 2, 3, \cdots$$

almost surely.

Proof. Since $L^{\mathbf{z}}_{K,s}$ is positive and $\|\gamma_t L^{\mathbf{z}}_{K,s}\|_{\mathfrak{L}(\mathcal{H}_K^n)} \leq 1$ for any $t \geq 1$, so $\|I - \gamma_t L^{\mathbf{z}}_{K,s}\|_{\mathfrak{L}(\mathcal{H}_K^n)} \leq 1$ a.s. for $t \geq 1$. We have from Lemma 5 that when $k \geq 2$,

$$\|\vec{f}^{\mathbf{z}}_k\|_{\mathcal{H}_K^n} \leq \sum_{l=1}^{k-1} \gamma_1 l^{-\tau} \frac{2\kappa M}{\sqrt{\mathrm{e}} s^{n+1}} \leq \frac{2\kappa M \gamma_1 \left((k-1)^{1-\tau} - \tau \right)}{\sqrt{\mathrm{e}} s^{n+1}(1-\tau)}$$

almost surely. □

Proof of Lemma 2. By definition, we get

$$\vec{f}_{k+1}^{\mathbf{z}} - \vec{f}_{k+1} = (1 - \gamma_k L_{K,s})(\vec{f}_k^{\mathbf{z}} - \vec{f}_k) + \gamma_k \chi_k,$$

where $\chi_k = (L_{K,s} - L_{K,s}^{\mathbf{z}})\vec{f}_k^{\mathbf{z}} + \vec{f}_{\rho,s}^{\mathbf{z}} - \vec{f}_{\rho,s}$. Since $\vec{f}_1^{\mathbf{z}} = \vec{f}_1 = 0$, we have by simple iteration:

$$\vec{f}_{k+1}^{\mathbf{z}} - \vec{f}_{k+1} = \sum_{j=1}^{k} \gamma_j \prod_{p=j+1}^{k} (1 - \gamma_p L_{K,s}) \chi_j,$$

so,

$$\|\vec{f}_{k+1}^{\mathbf{z}} - \vec{f}_{k+1}\|_{\mathcal{H}_K^n} \leq \sum_{j=1}^{k} \gamma_j \|\chi_j\|_{\mathcal{H}_K^n}$$

with confidence $1 - 2\delta$. Hence

$$\begin{aligned}
\|\vec{f}_{k+1}^{\mathbf{z}} - \vec{f}_{k+1}\|_{\mathcal{H}_K^n} &\leq \sum_{j=1}^{k} \gamma_j \left(\frac{34\kappa^2 \sqrt{n} \log \frac{2}{\delta}}{\mathrm{e}\sqrt{m} s^n} \cdot \frac{2\kappa M \gamma_1 j^{1-\tau}}{\sqrt{\mathrm{e}} s^{n+1}(1-\tau)} + \frac{34 M \kappa \log \frac{2}{\delta}}{\sqrt{\mathrm{e}m} s^{n+1}} \right) \\
\leq & \frac{34\sqrt{n} M \log \frac{2}{\delta}}{\kappa \mathrm{e}^{3/2} \sqrt{m} s (1-\tau)^2} (k+1)^{2-2\tau} + \frac{34 M \log \frac{2}{\delta}}{\kappa \sqrt{\mathrm{e}m} s (1-\tau)} (k+1)^{1-\tau} \\
\leq & \frac{34 M \log \frac{2}{\delta}}{\kappa \sqrt{\mathrm{e}m} s (1-\tau)^2} (k+1)^{2-2\tau} \left(\frac{\sqrt{n}}{\mathrm{e}} + 1 \right) = \frac{C_3 (k+1)^{2-2\tau}}{\sqrt{m} s (1-\tau)^2} \log \frac{2}{\delta}.
\end{aligned}$$

□

Let $\mathfrak{M}(H)$ denote the class of all the sequence $f = (f_0, f_1, \cdots)$ of Bochner-integrable random variables with values in the separable Hilbert space H such that $f_0 = 0$ and f is a martingale. Pinelis proved the following result ([11], special case with $D = 1$ of Theorem 3.2).

Lemma 7. *Let H be a separable Hilbert space, $f \in \mathfrak{M}(H)$ and f be adapted to a non-decreasing sequence $\{F_j\}_{j=0}^{\infty}$ of sub-σ-fields of the Borel set $\mathfrak{B}$ on the probability space Ω. Suppose $\lambda > 0$ satisfies that $\mathbb{E}\mathrm{e}^{\lambda \|d_j\|} < +\infty$ for $j = 1, 2, \cdots$, where $d_j = f_j - f_{j-1}$. Then for all $r \geq 0$,*

$$\mathrm{Prob}\left\{ \sup_j \|f_j\| \geq r \right\} \leq 2 \exp\left\{ -\lambda r + \left\| \sum_{j=1}^{\infty} e_j \right\|_{\infty} \right\},$$

where $e_j := \mathbb{E}\left\{ \mathrm{e}^{\lambda \|d_j\|} - 1 - \lambda \|d_j\| \,\middle|\, F_{j-1} \right\} \geq 0$, a.e.. □

Lemma 7 directly implies

Lemma 8. *For a finite martingale $f = (f_0, \cdots, f_m)$, $f_0 = 0$, with the same settings as Lemma 7, one has*

$$\mathrm{Prob}\left\{\max_{1\le j\le m}\|f_j\| \ge r\right\} \le 2\exp\left\{-\lambda r + m(\mathrm{e}^{\lambda\Delta} - 1 - \lambda\Delta)\right\},$$

where $\Delta \ge \max_{1\le j\le m}\|d_j\|_\infty$. □

One can obtain the following corollary directly by modifying Pinelis' proof [11] to Lemma 7 by a few lines. Probability inequalities of the similar type are also proved in [11].

Corollary 1. *For a finite martingale $f = (f_0, \cdots, f_m)$, $f_0 = 0$, with the same settings as Lemma 7, for any $\Delta \ge 0$, one has*

$$\mathrm{Prob}\left(\max_{1\le j\le m}\|f_j\| \ge r,\ \max_{1\le i\le m}\|d_i\| \le \Delta\right) \le 2\exp\left\{-\lambda r + m\left(\mathrm{e}^{\lambda\Delta} - 1 - \lambda\Delta\right)\right\}.$$

Proof. As was done in [11], we build a positive super-martingale

$$G_0 = 1,\ G_j = \cosh(\lambda\|f_j\|) \Big/ \prod_{i=1}^{j}(1+e_i)\ ,\ j = 1, \cdots, m.$$

We denote $J := \min\{j : \|f_j\| \ge r\}$ if it exists. Since f is a finite martingale, one has $J \le m$. Thus

$$\begin{aligned}
& \mathrm{Prob}\left(\max_{1\le j\le m}\|f_j\| \ge r \,\middle|\, \max_{1\le i\le m}\|d_i\| \le \Delta\right) \\
\le\ & \mathrm{Prob}\left(G_J \ge \cosh(\lambda r) \Big/ \prod_{j=1}^{m}(1+e_j) \,\middle|\, \max_{1\le i\le m}\|d_i\| \le \Delta\right) \\
\le\ & \mathrm{Prob}\left(G_J \ge \frac{\mathrm{e}^{\lambda r}}{2} \Big/ \prod_{j=1}^{m}\left(1+\mathrm{e}^{\lambda\Delta} - 1 - \lambda\Delta\right) \,\middle|\, \max_{1\le i\le m}\|d_i\| \le \Delta\right) \\
\le\ & \frac{2\mathbb{E}\left(G_J \,|\max_{1\le i\le m}\|d_i\| \le \Delta\right)}{\mathrm{e}^{\lambda r}}\left(1 + \left(\mathrm{e}^{\lambda\Delta} - 1 - \lambda\Delta\right)\right)^m,
\end{aligned}$$

where Chebyshev's inequality is used in the last step. Since G_J is non-negative, one has

$$\mathrm{Prob}\left(\max_{1\le i\le m}\|d_i\| \le \Delta\right)\cdot\mathbb{E}\left(G_J \,\middle|\, \max_{1\le i\le m}\|d_i\| \le \Delta\right) \le \mathbb{E}G_J \le \mathbb{E}G_0 = 1.$$

Also, since for all $t \geq 0$, $\mathrm{e}^t - 1 - t \geq 0$, hence for all $p \geq 0$, $1 + p \leq \mathrm{e}^p$, so we have

$$\begin{aligned}\left(1 + \mathrm{e}^{\lambda\Delta} - 1 - \lambda\Delta\right)^m &= \exp\left(m \log(1 + (\mathrm{e}^{\lambda\Delta} - 1 - \lambda\Delta))\right) \\ &\leq \exp\left(m(\mathrm{e}^{\lambda\Delta} - 1 - \lambda\Delta)\right),\end{aligned}$$

which implies the conclusion. □

In the large dimension, small sample problem, a primary observation is that the probability of any two sample points be very close should be very small. To formulate the fact precisely, for $\mathbf{x} = \{x_i\}_{i=1}^m$ drawn i.i.d. from ρ_X, we give the following

Lemma 9. *For any $\delta \in (0, 1)$, with confidence $1 - \delta$, we have*

$$\varepsilon_{\mathbf{z}} \geq \left(\frac{\delta n \Gamma(n/2)}{\pi^{n/2} c_p m^2}\right)^{1/n}.$$

Proof. Since $\mathbf{x}$ is i.i.d. drawn, for any $\varepsilon_0 > 0$,

$$\begin{aligned}&\mathrm{Prob}(\varepsilon^{\mathbf{z}} < \varepsilon_0) \leq \sum_{1 \leq i < j \leq m} \mathrm{Prob}(|x_i - x_j| < \varepsilon_0) = \binom{m}{2} \mathrm{Prob}(|x_1 - x_2| < \varepsilon_0) \\ \leq\ & \frac{m^2}{2} \int_X \mathrm{d}\rho_X(x_1) \int_{B(x_1, \varepsilon_0) \cap X} \mathrm{d}\rho_X(x_2) \leq \frac{m^2}{2} \int_X \mathrm{d}\rho_X(x_1) \int_{B(x_1, \varepsilon_0)} c_p \, \mathrm{d}x_2 \\ =\ & \frac{m^2 c_p}{2} \int_X \frac{2\pi^{n/2} \varepsilon_0^n}{n\Gamma(n/2)} \mathrm{d}\rho_X(x_1) = \frac{\pi^{n/2} c_p \varepsilon_0^n m^2}{n\Gamma(n/2)},\end{aligned}$$

which implies the result. □

Lemma 10. *Let $n \geq 23$, with confidence $1 - 3\delta$ for $\delta \in (0, 1/3)$, we have*

$$\|L_{K,s}^{\mathbf{z}} - L_{K,s}\|_{\mathrm{HS}} \leq \frac{5\sqrt{n}\,\kappa^2\, c_p\, J_2}{\sqrt{m}} \log \frac{2\mathrm{e}}{\delta} \tag{4.8}$$

$$\|\vec{f}_{\rho,s}^{\mathbf{z}} - \vec{f}_{\rho,s}\|_{\mathcal{H}_K^n} \leq \frac{26\kappa M c_p J_2}{C_4} m^{\frac{2\alpha - 9}{2(11 - 2\alpha)}} \left(2\sqrt{\log m} + 1\right) \delta^{-\frac{1}{n}} \left(\log \frac{2\mathrm{e}}{\delta}\right)^{\frac{3}{2}} \tag{4.9}$$

where s and C_4 are set coherent with Lemma 3.

Proof. Consider

$$\frac{\partial}{\partial t}\left(\frac{t^2}{s^{n+2}} \mathrm{e}^{-t^2/2s^2}\right) = \frac{1}{s^{n+2}} \left(2t - \frac{t^3}{s^2}\right) \mathrm{e}^{-t^2/2s^2} \tag{4.10}$$

$$\frac{\partial}{\partial s}\left(\frac{t^2}{s^{n+2}} \mathrm{e}^{-t^2/2s^2}\right) = t^2 \left(-\frac{n+2}{s^{n+3}} + \frac{t^2}{s^{n+5}}\right) \mathrm{e}^{-t^2/2s^2}, \tag{4.11}$$

we see that when $0 < s \leq \frac{\varepsilon_{\mathbf{z}}}{\sqrt{n+2}}$, and $t \geq \varepsilon_{\mathbf{z}}$, the function $\frac{t^2}{s^{n+2}} \mathrm{e}^{-t^2/2s^2}$ is increasing w.r.t. s, and decreasing w.r.t. t, so by (4.4) we have

$$\|L_{K,s}^{\mathbf{z}}\|_{\mathrm{HS}} \leq \frac{\kappa^2\sqrt{n}}{m^2} \sum_{i,j=1}^{m} w_{ij}^{(s)} |x_i - x_j|^2 \leq \frac{\kappa^2\sqrt{n}(m-1)\varepsilon_{\mathbf{z}}^2}{ms^{n+2}} \exp\left\{-\frac{\varepsilon_{\mathbf{z}}^2}{2s^2}\right\}.$$

Since

$$0 < s \leq \frac{\varepsilon_{\mathbf{z}}}{\frac{2(n+2)}{\mathrm{e}} + \sqrt{2|\log(\varepsilon_{\mathbf{z}}^n c_p J_2)|}},$$

we have

$$\begin{aligned}\left(\frac{\varepsilon_{\mathbf{z}}}{s} - \frac{n+2}{\mathrm{e}}\right)^2 &\geq \left(\frac{n+2}{\mathrm{e}} + \sqrt{2|\log(\varepsilon_{\mathbf{z}}^n c_p J_2)|}\right)^2 \\ &\geq \left(\frac{n+2}{\mathrm{e}}\right)^2 - 2\log(\varepsilon_{\mathbf{z}}^n c_p J_2),\end{aligned}$$

hence

$$-\frac{\varepsilon_{\mathbf{z}}^2}{2s^2} + \frac{(n+2)\varepsilon_{\mathbf{z}}}{\mathrm{e}s} \leq \log(\varepsilon_{\mathbf{z}}^n c_p J_2).$$

Because $\log t \leq \frac{t}{\mathrm{e}}$ for any $t > 0$, we have

$$-\frac{\varepsilon_{\mathbf{z}}^2}{2s^2} + (n+2)\log\frac{\varepsilon_{\mathbf{z}}}{s} \leq \log(\varepsilon_{\mathbf{z}}^n c_p J_2),$$

that is,

$$\frac{\varepsilon_{\mathbf{z}}^2}{s^{n+2}} \mathrm{e}^{-\frac{\varepsilon_{\mathbf{z}}^2}{2s^2}} \leq c_p J_2, \quad \text{a.s.}, \tag{4.12}$$

so,

$$\|L_{K,s}^{\mathbf{z}}\|_{\mathrm{HS}} \leq \kappa^2\sqrt{n} c_p J_2 \tag{4.13}$$

almost surely. Owing to the continuity of $L_{K,s}^{\mathbf{z}}$ with respect to $z_1, \cdots, z_m$, $L_{K,s}^{\mathbf{z}}$ is a Bochner integrable random variable.

We define a sequence $f = (f_0, f_1, \cdots, f_m)$ with $f_0 = 0$ and

$$f_i = \mathbb{E}\left\{ L_{K,s}^{\mathbf{z}} - \frac{m-1}{m} L_{K,s} \middle| z_1, \cdots, z_i \right\}, \ \ i = 1, \cdots, m.$$

Then f is a martingale. We define $d_j = f_j - f_{j-1}$ for $1 \le j \le m$. From (4.5) and (4.13), we see that f_j's are uniformly bounded, so are d_j's, $j = 0, 1, \cdots, m$, thus $\mathbb{E}\mathrm{e}^{\lambda\|d_j\|_{\mathrm{HS}}} < +\infty$ for any $1 \le j \le m$ and $\lambda \ge 0$.

We have

$$d_j \;=\; \mathbb{E}\left\{ L^{\mathbf{z}}_{K,s} - \mathbb{E}_{z_j} L^{\mathbf{z}}_{K,s} \middle| z_1, \cdots, z_j \right\}.$$

Now,

$$\begin{aligned}
& L^{\mathbf{z}}_{K,s} - \mathbb{E}_{z_j} L^{\mathbf{z}}_{K,s} \\
= \;& \frac{1}{m^2} \sum_{i=1}^{m} w_{ij}^{(s)} (x_j - x_i)(x_j - x_i)^T (A_{x_j} + A_{x_i}) \\
& - \frac{1}{m^2} \sum_{i=1, i\neq j}^{m} \int_X w(x, x_i)(x - x_i)(x - x_i)^T (A_x + A_{x_i}) \, \mathrm{d}\rho_X(x) \\
=: \;& W_1 - W_2,
\end{aligned}$$

and

$$\begin{aligned}
\|W_2\|_{\mathrm{HS}} \;&\le\; \frac{2\sqrt{n}\kappa^2}{m^2} \sum_{i=1, i\neq j}^{m} \int_X \frac{1}{s^{n+2}} \exp\left\{ \frac{-|x - x_i|^2}{2s^2} \right\} |x - x_i|^2 p(x) \, \mathrm{d}x \\
&\le\; \frac{2\kappa^2\sqrt{n}}{m} c_p J_2.
\end{aligned}$$

Following from (4.10), (4.11), and (4.12),

$$\begin{aligned}
\|W_1\|_{\mathrm{HS}} \;&\le\; \frac{2\sqrt{n}\kappa^2}{m^2} \sum_{i=1}^{m} w(x_i, x_j)|x_i - x_j|^2 \\
&\le\; \frac{2\sqrt{n}\kappa^2 \varepsilon_{\mathbf{z}}^2}{m s^{n+2}} \exp\left\{ -\frac{\varepsilon_{\mathbf{z}}^2}{2s^2} \right\} \le \frac{2\sqrt{n}\kappa^2 c_p J_2}{m}.
\end{aligned}$$

So,

$$\|d_j\|_{\mathrm{HS}} \le \|W_1\|_{\mathrm{HS}} + \|W_2\|_{\mathrm{HS}} \le \frac{4\sqrt{n}\kappa^2 c_p J_2}{m}$$

almost surely.

Using Lemma 8 by taking $\Delta = 4\sqrt{n}\kappa^2 c_p J_2/m$ and $\lambda = \frac{1}{\Delta\sqrt{m}} \le \frac{1}{\Delta}$ which implies $\mathrm{e}^{\lambda\Delta} - 1 - \lambda\Delta \le (\lambda\Delta)^2 = \frac{1}{m}$, we have for any $r_1 > 0$,

$$\mathrm{Prob}\left\{ \max_{1\le j\le m} \|f_j\|_{\mathrm{HS}} \ge r_1 \right\} \le 2\exp\left\{ -\frac{r_1}{\Delta\sqrt{m}} + 1 \right\}. \tag{4.14}$$

Put $\delta = 2\exp\left\{-\frac{r_1}{\Delta\sqrt{m}}+1\right\}$, we get $r_1 = \frac{4\sqrt{n}\kappa^2 c_p J_2}{\sqrt{m}}\log\frac{2\mathrm{e}}{\delta}$, so, with confidence $1-\delta$,

$$\left\| L_{K,s}^{\mathbf{z}} - \frac{m-1}{m} L_{K,s} \right\|_{\mathrm{HS}} \le \max_{1\le j\le m} \|f_j\|_{\mathrm{HS}} \le \frac{4\sqrt{n}\kappa^2 c_p J_2}{\sqrt{m}} \log\frac{2\mathrm{e}}{\delta}, \qquad (4.15)$$

which, combined with (4.5), proves (4.8).

We let now $f_i' := \mathbb{E}\left\{ \vec{f}_{\rho,s}^{\mathbf{z}} - \frac{m-1}{m}\vec{f}_{\rho,s} \middle| z_1, \cdots, z_i \right\}$, $i = 1, \cdots, m$, and $f_0' = 0$. $\{f_i'\}$ also forms a finite martingale with each random variable taking value in $\mathcal{H}_K^n$. We define $d_j' = f_j' - f_{j-1}'$ for $1 \le j \le m$. Similarly,

$$d_j' = \mathbb{E}\{\vec{f}_{\rho,s}^{\mathbf{z}} - \mathbb{E}_{z_j}\vec{f}_{\rho,s}^{\mathbf{z}} | z_1, \cdots, z_j\}.$$

Now

$$\begin{aligned} & \vec{f}_{\rho,s}^{\mathbf{z}} - \mathbb{E}_{z_j}\vec{f}_{\rho,s}^{\mathbf{z}} \\ = \quad & \frac{1}{m^2}\sum_{i=1}^m w_{ij}(y_j - y_i)(x_j - x_i)(K_{x_j} + K_{x_i}) \\ & -\frac{1}{m^2}\sum_{i=1, i\ne j}^m \int_X w(x, x_i)(f_\rho(x) - y_i)(x - x_i)(K_x + K_{x_i})\,\mathrm{d}\rho_X(x) \\ =: \quad & W_1' - W_2'. \end{aligned}$$

Since $|y_i| \le M$ a.s. for $i = 1, \cdots, m$, we have

$$\|W_1'\|_{\mathcal{H}_K^n} \le \frac{1}{m^2}\sum_{i=1}^m 4\kappa M \frac{|x_i - x_j|}{s^{n+2}} \mathrm{e}^{-|x_i - x_j|^2/2s^2}.$$

Thanks to

$$\frac{\partial}{\partial t}\left(\frac{t}{s^{n+2}}\mathrm{e}^{-\frac{t^2}{2s^2}}\right) = \frac{1}{s^{n+2}}\left(1 - \frac{t^2}{s^2}\right)\mathrm{e}^{-\frac{t^2}{2s^2}},$$

we see that when $s \le \frac{t}{\sqrt{n+2}} < t$, the function $\frac{t}{s^{n+2}}\mathrm{e}^{-t^2/2s^2}$ is decreasing w.r.t. t, so

$$\|W_1'\|_{\mathcal{H}_K^n} \le \frac{4M\kappa\varepsilon_{\mathbf{z}}}{ms^{n+2}} \exp\left\{-\frac{\varepsilon_{\mathbf{z}}^2}{2s^2}\right\} \le \frac{4M\kappa c_p J_2}{m\varepsilon_{\mathbf{z}}},$$

where the second inequality follows from (4.12). The next inequality is derived easily from the fact $J_1 \le J_2$ as

$$\|W_2'\|_{\mathcal{H}_K^n} \le \frac{1}{m} 4\kappa M c_p J_1 s^{-1} \le \frac{4\kappa M c_p J_2}{ms}.$$

So we get

$$\|d'_j\|_{\mathcal{H}_K^n} \leq \frac{4\kappa M c_p J_2}{m}\left(\frac{1}{\varepsilon_{\mathbf{z}}}+\frac{1}{s}\right) \leq \frac{8\kappa M c_p J_2}{ms}, \quad \text{a.s.},$$

where the second inequality comes from $s \leq \frac{\varepsilon_{\mathbf{z}}}{\sqrt{n+2}} \leq \varepsilon_{\mathbf{z}}$.

By definition $s_0^{\mathbf{z}} \leq 1$, on the other hand, by Lemma 9 we have with confidence $1-\delta$,

$$\varepsilon_{\mathbf{z}} \geq \left(\frac{\delta n \Gamma(n/2)}{\sqrt{\pi^n} c_p m^2}\right)^{1/n}, \tag{4.16}$$

which implies

$$|\log \varepsilon_{\mathbf{z}}| \leq |\log \mathrm{Diam}(X)| + \frac{2}{n}\log m + \frac{1}{n}\left|\log\left(\frac{\delta n \Gamma(n/2)}{\sqrt{\pi^n} c_p}\right)\right|.$$

Therefore, when (4.16) holds, we have

$$\frac{\varepsilon_{\mathbf{z}}}{\frac{2(n+2)}{\mathrm{e}} + \sqrt{2|\log(\varepsilon_{\mathbf{z}}^n c_p J_2)|}} \geq \frac{C_X m^{-2/n}\delta^{1/n}}{\left(2\sqrt{\log m}+1\right)\left(\sqrt{2\log\frac{1}{\delta}}+1\right)},$$

where

$$C_X := \frac{\left(\frac{n\Gamma(n/2)}{\sqrt{\pi^n}c_p}\right)^{1/n}}{\frac{2(n+2)}{\mathrm{e}} + \sqrt{2|\log(c_p J_2)|} + \sqrt{2n|\log \mathrm{Diam}(X)| + 2\left|\log\left(\frac{n\Gamma(n/2)}{\sqrt{\pi^n}c_p}\right)\right|}},$$

which depends only on (X, ρ_X). Since $\delta \in (0, 1/3)$, then $\delta \leq 1/\sqrt{\mathrm{e}}$, which implies

$$\sqrt{2\log\frac{1}{\delta}} \geq 1,$$

so by definition we have with confidence $1-\delta$,

$$s_0^{\mathbf{z}} \geq C_4 \frac{m^{-2/n}\delta^{1/n}}{3\left(2\sqrt{\log m}+1\right)\sqrt{\log\frac{1}{\delta}}}, \tag{4.17}$$

where

$$C_4 = \min\left\{1, \frac{R}{3\sqrt{n+4}}, C_X\right\}$$

depends also only on (X, ρ_X).

Therefore, we have with confidence $1-\delta$, $\|d'_j\|_{\mathcal{H}_K^n} \leq \Delta'$, where

$$\Delta' = \frac{24\kappa M c_p J_2}{C_4 \delta^{1/n}} m^{\frac{2\alpha-10}{11-2\alpha}} \left(2\sqrt{\log m}+1\right) \sqrt{\log \frac{1}{\delta}}.$$

We take $\lambda' = \frac{1}{\Delta'\sqrt{m}}$ which implies $\mathrm{e}^{\lambda'\Delta'} - 1 - \lambda'\Delta' \leq \frac{1}{m}$. So, for any $r_2 \geq 0$, by Corollary 1,

$$\begin{aligned} & \mathrm{Prob}\left\{\max_{1\leq j\leq m} \|f'_j\|_{\mathcal{H}_K^n} \geq r_2\right\} \\ \leq\ & \mathrm{Prob}\left\{\max_{1\leq j\leq m} \|f'_j\|_{\mathcal{H}_K^n} \geq r_2,\ \max_{1\leq j\leq m} \|d'_j\|_{\mathcal{H}_K^n} \leq \Delta'\right\} + \delta \\ \leq\ & \delta + 2\exp\left\{-\frac{r_2}{\Delta'\sqrt{m}} + 1\right\}. \end{aligned}$$

Put $\delta = 2\exp\left\{-\frac{r_2}{\Delta'\sqrt{m}} + 1\right\}$, we have

$$r_2 = \Delta'\sqrt{m}\log\frac{2\mathrm{e}}{\delta} \tag{4.18}$$

$$\leq \frac{24\kappa M c_p J_2}{C_4 \delta^{1/n}} m^{\frac{-9+2\alpha}{2(11-2\alpha)}} \left(2\sqrt{\log m}+1\right) \left(\log\frac{2\mathrm{e}}{\delta}\right)^{\frac{3}{2}}, \tag{4.19}$$

thus with confidence $1-2\delta$,

$$\left\| \vec{f}^{\mathbf{z}}_{\rho,s} - \frac{m-1}{m} \vec{f}_{\rho,s} \right\|_{\mathcal{H}_K^n} \leq \max_{1\leq j\leq m} \|f'_j\|_{\mathcal{H}_K^n} \leq r_2, \tag{4.20}$$

which, combined with (4.19) and the estimation $\frac{1}{m}\|\vec{f}_{\rho,s}\|_{\mathcal{H}_K^n} \leq \frac{2\kappa M c_p J_2}{m}$, proves (4.9). The proof is thus completed. □

Corollary 2. *when (4.20) holds,*

$$\|\vec{f}^{\mathbf{z}}_{\rho,s}\|_{\mathcal{H}_K^n} \leq \frac{26\kappa M c_p J_2}{C_4 \delta^{1/n}} \left(2\sqrt{\log m}+1\right) \left(\log\frac{2\mathrm{e}}{\delta}\right)^{\frac{3}{2}}.$$

Proof. Direct computing verifies the result. □

Lemma 11. *For $\vec{f}^{\mathbf{z}}_k$ recurrently defined in (1.2) and $k \geq 2$, we have*

$$\vec{f}^{\mathbf{z}}_k = \sum_{l=1}^{k-1} \gamma_l \prod_{p=l+1}^{k-1} \left(I - \gamma_p L^{\mathbf{z}}_{K,s}\right) \vec{f}^{\mathbf{z}}_{\rho,s}, \tag{4.21}$$

where we denote $\prod_{p=k}^{k-1}\left(1-\gamma_p L_{K,s}^{\mathbf{z}}\right) := I$ *for saving the notations. Moreover, when (4.9) holds true, setting* $0<\gamma_1\leq\left(\sqrt{n}\kappa^2 c_p J_2\right)^{-1}$*, we have*

$$\|\vec{f}_k^{\mathbf{z}}\|_{\mathcal{H}_K^n}\leq\frac{26M}{\kappa\sqrt{n}C_4\delta^{1/n}}\left(2\sqrt{\log m}+1\right)\left(\log\frac{2\mathrm{e}}{\delta}\right)^{\frac{3}{2}}\frac{(k-1)^{1-\tau}}{1-\tau}.$$

Proof. (4.21) could be verified directly by computing. From (4.13), we have $1-\gamma_p\|L_{K,s}^{\mathbf{z}}\|_{\mathfrak{L}(\mathcal{H}_K^n)}\geq 0$. Since $L_{K,s}^{\mathbf{z}}$ is positive, $\|1-\gamma_p L_{K,s}^{\mathbf{z}}\|_{\mathfrak{L}(\mathcal{H}_K^n)}\leq 1$. So when (4.9) holds true, for any $k\geq 2$,

$$\begin{aligned}\|\vec{f}_k^{\mathbf{z}}\| &\leq \sum_{l=1}^{k-1}\gamma_l\|\vec{f}_{\rho,s}^{\mathbf{z}}\|_{\mathcal{H}_K^n}\\ &\leq \frac{26M}{\kappa\sqrt{n}C_4\delta^{1/n}}\left(2\sqrt{\log m}+1\right)\left(\log\frac{2\mathrm{e}}{\delta}\right)^{\frac{3}{2}}\frac{(k-1)^{1-\tau}}{1-\tau}.\end{aligned}$$

□

Proof of Lemma 3. By definition, we get

$$\vec{f}_{k+1}^{\mathbf{z}}-\vec{f}_{k+1}=(1-\gamma_k L_{K,s})(\vec{f}_k^{\mathbf{z}}-\vec{f}_k)+\gamma_k\chi_k,$$

where $\chi_k=(L_{K,s}-L_{K,s}^{\mathbf{z}})\vec{f}_k^{\mathbf{z}}+\vec{f}_{\rho,s}^{\mathbf{z}}-\vec{f}_{\rho,s}$. Since $\vec{f}_1^{\mathbf{z}}=\vec{f}_1=0$, we have by simple iteration:

$$\vec{f}_{k+1}^{\mathbf{z}}-\vec{f}_{k+1}=\sum_{j=1}^{k}\gamma_j\left(\prod_{q=j+1}^{k}(1-\gamma_q L_{K,s})\right)\chi_j.$$

Since $L_{K,s}\in\mathfrak{L}(\mathcal{H}_K^n)$ is positive,

$$\begin{aligned}&\|L_{K,s}\|_{\mathfrak{L}(\mathcal{H}_K^n)}=\sup_{\vec{g}\in\mathcal{H}_K^n,\|\vec{g}\|_{\mathcal{H}_K^n}=1}\langle L_{K,s}\vec{g},\vec{g}\rangle_{\mathcal{H}_K^n}\\ =&\sup_{\vec{g}\in\mathcal{H}_K^n,\|\vec{g}\|_{\mathcal{H}_K^n}=1}\int_X\int_X w(x,u)\left((u-x)^T\vec{g}(x)\right)^2\,\mathrm{d}\rho_X(u)\,\mathrm{d}\rho_X(x)\leq\kappa^2 c_p J_2,\end{aligned}$$

thus $1-\gamma_q\|L_{K,s}\|_{\mathfrak{L}(\mathcal{H}_K^n)}\geq 0$, so for any $q\geq 1$, $\|1-\gamma_q L_{K,s}\|_{\mathfrak{L}(\mathcal{H}_K^n)}\leq 1$. We have

$$\|\vec{f}_{k+1}^{\mathbf{z}}-\vec{f}_{k+1}\|_{\mathcal{H}_K^n}\leq\sum_{j=1}^{k}\gamma_j\|\chi_j\|_{\mathcal{H}_K^n}.$$

Since (4.8), (4.9) and Lemma 11 imply

$$\begin{aligned}\|\chi_j\|_{\mathcal{H}_K^n} &\leq \|L_{K,s}^{\mathbf{z}} - L_{K,s}\|_{\mathfrak{L}(\mathcal{H}_K^n)}\|\vec{f}_j^{\mathbf{z}}\|_{\mathcal{H}_K^n} + \|\vec{f}_{\rho,s}^{\mathbf{z}} - \vec{f}_{\rho,s}\|_{\mathcal{H}_K^n} \\ &\leq \frac{5\sqrt{n}\kappa^2 c_p J_2}{\sqrt{m}}\left(\log\frac{2\mathrm{e}}{\delta}\right)^{\frac{5}{2}}\frac{26M(j-1)^{1-\tau}}{\kappa\sqrt{n}C_4\delta^{1/n}(1-\tau)}\left(2\sqrt{\log m}+1\right) \\ &\quad + \frac{26\kappa M c_p J_2}{C_4\delta^{1/n}} m^{\frac{2\alpha-9}{2(11-2\alpha)}}\left(2\sqrt{\log m}+1\right)\left(\log\frac{2\mathrm{e}}{\delta}\right)^{\frac{3}{2}},\end{aligned}$$

for $j = 1, 2, \cdots$, we have with confidence $1 - 3\delta$,

$$\begin{aligned}\|\vec{f}_{k+1}^{\mathbf{z}} - \vec{f}_{k+1}\|_{\mathcal{H}_K^n} &\leq \frac{65M(k+1)^{2-2\tau}\left(2\sqrt{\log m}+1\right)}{\kappa\sqrt{mn}C_4\delta^{1/n}(1-\tau)^2}\left(\log\frac{2\mathrm{e}}{\delta}\right)^{\frac{5}{2}} \\ &\quad + \frac{26M\left(2\sqrt{\log m}+1\right)(k+1)^{1-\tau}}{\kappa\sqrt{n}C_4\delta^{1/n}(1-\tau)} m^{\frac{2\alpha-9}{2(11-2\alpha)}}\left(\log\frac{2\mathrm{e}}{\delta}\right)^{\frac{3}{2}},\end{aligned}$$

which implies the result. □

5. Approximation Error

We put here the approximation error estimation first.

Theorem 4. *For the global iteration and the step size $\gamma_t = \gamma_1 t^{-\tau}$ with $0 < \gamma_1 \leq \left(\kappa^2 c_p J_2\right)^{-1}$ and $0 \leq \tau < 1$, if $k \geq 1$, one has*

$$\begin{aligned}\|\vec{f}_{k+1} - \nabla f_\rho\|_\rho \leq{}& \frac{\|L_K^{-1}\nabla f_\rho\|_\rho(1-\tau)}{\mathrm{e}ws^\alpha\gamma_1(1-2^{\tau-1})(k+1)^{1-\tau}} \\ &+ C_5\gamma_1\kappa^2 s^{3/2} + \frac{6C_5}{\mathrm{e}w}s^{\frac{3}{2}-\alpha}\log\frac{k+1}{1-\tau},\end{aligned}$$

with w, s, and C_5 set in (3.3), (5.5), and (5.6) respectively.

In the analysis of this section, we assume that the regression function f_ρ has the following regularity

$$M_\nu := \operatorname*{ess\,sup}_{x\in X}\left(\sum_{1\leq i_1,\cdots,i_\nu\leq n}\left(\frac{\partial^\nu f_\rho(x)}{\partial x^{i_1}\cdots\partial x^{i_\nu}}\right)^2\right)^{1/2} < +\infty \tag{5.1}$$

with $\nu = 2, 3$. We assume for the density function $p(x)$,

$$M_p := \operatorname*{ess\,sup}_{x\in X}|\nabla p(x)| = \operatorname*{ess\,sup}_{x\in X}\left(\sum_{i=1}^n\left(\frac{\partial p(x)}{\partial x^i}\right)^2\right)^{1/2} < +\infty. \tag{5.2}$$

We define

$$\psi(r) = \rho_X(\{x \in X : \mathrm{dist}(x, \partial X) \le r\}), \tag{5.3}$$

then $\forall r \le 0, \psi(r) = 0$, and $\forall r \ge \mathrm{Diam}(X)/2$, $\psi(r) = 1$, where $\mathrm{Diam}(X) := \sup_{x,y\in X} |x - y|$. $\psi(r)$ is an increasing function and so it is differentiable a.e.. We assume that ψ is absolutely continuous with its derivative $\psi'(r)$ bounded:

$$|\psi'(r)| \le M_{\psi'} < +\infty \tag{5.4}$$

for a.e. $r \in \mathbb{R}$. For the weight parameter s, we require during this section that

$$0 < s \le \min\left\{1, \frac{R}{3\sqrt{n+4}}\right\} \tag{5.5}$$

with R set as were in Theorem 3. Denote

$$\vec{\zeta}(x) := \int_X w(x,u)(f_\rho(u) - f_\rho(x))(u - x)\,\mathrm{d}\rho_X(u),$$

then $L_K\vec{\zeta} = \vec{f}_{\rho,s}$, and we have

Lemma 12. *With regularity assumptions (5.1), (5.2) and (5.4) being satisfied, one has*

$$\|\vec{\zeta} - \Gamma_s \nabla f_\rho\|_\rho \le C_5 s^{3/2},$$

where

$$\begin{aligned} C_5 &= \frac{\Gamma((n+3)/2)}{\Gamma(n/2)} M_2 c_p 2^{\frac{n+1}{2}} \sqrt{(n+3)\pi^n M_{\psi'}} \\ &\quad + \frac{1}{6} M_3 c_p J_4 + \frac{1}{2} M_2 M_p J_4 + \frac{1}{6} M_3 M_p J_5. \end{aligned} \tag{5.6}$$

Proof. For any $x \in X$, we write $r(x) := \mathrm{dist}(x, \partial X)$, then

$$\begin{aligned} & \left|\vec{\zeta}(x) - \Gamma_s(x)\nabla f_\rho(x)\right| \\ \le\ & \left|\int_{B(x,r(x))} w(x,u)(f_\rho(u) - f_\rho(x) - \nabla f_\rho(x)^T(u-x))(u-x)\,\mathrm{d}\rho_X(u)\right| \\ & + \int_{X\setminus B(x,r(x))} w(x,u)\frac{M_2}{2}|u-x|^3 p(u)\,\mathrm{d}u \\ =:\ & I_1 + I_2, \end{aligned}$$

where the inequality holds because

$$f_\rho(u) - f_\rho(x) - \nabla f_\rho(x)^T(u-x) = \frac{1}{2}(u-x)^T \mathrm{Hess} f_\rho(x+\theta_x(u)(u-x))(u-x),$$

with $0 < \theta_x(u) < 1$.

Doing one step further the expansion:

$$\begin{aligned} & f_\rho(u) - f_\rho(x) - \nabla f_\rho(x)^T(u-x) \\ = \; & \frac{1}{2}(u-x)^T \mathrm{Hess} f_\rho(x)(u-x) \\ & + \frac{1}{6}\sum_{i,j,k=1}^{n} \frac{\partial^3 f_\rho(x+\widetilde{\theta}_x(u)(u-x))}{\partial x^i \partial x^j \partial x^k}(u^i - x^i)(u^j - x^j)(u^k - x^k), \end{aligned}$$

and

$$p(u) = p(x) + \nabla p(x+\mu_x(u)(u-x))^T(u-x),$$

where $\widetilde{\theta}_x(u), \mu_x(u) \in (0,1)$, we have

$$\begin{aligned} I_1 \leq & \left| \int_{B(x,r(x))} w(x,u)\frac{1}{2}\left((u-x)^T \mathrm{Hess} f_\rho(x)(u-x)\right)(u-x)p(x)\,\mathrm{d}u \right| \\ & + \int_{B(x,r(x))} w(x,u)|u-x| \left(\frac{1}{6}|u-x|^3 M_3 p(x) + \frac{1}{2}|u-x|^3 M_2 M_p \right. \\ & \qquad \left. + \frac{1}{6}|u-x|^4 M_3 M_p \right) \mathrm{d}u. \end{aligned}$$

By a change of variable $v = \frac{u-x}{s}$, we see that

$$\begin{aligned} I_1 & \leq 0 + \int_{B(0,r(x)/s)} s^2 \mathrm{e}^{-|v|^2/2}|v|^4 \left(\frac{1}{6}M_3 p(x) + \frac{1}{2}M_2 M_p + \frac{s}{6}|v| M_3 M_p \right) \mathrm{d}v \\ & \leq s^2 \left(\frac{1}{6}M_3 c_p + \frac{1}{2}M_2 M_p \right) J_4 + \frac{s^3}{6} M_3 M_p J_5, \end{aligned}$$

since $s \leq 1$, we have

$$\|I_1\|_\rho \leq s^2 \left(\frac{1}{6}M_3 c_p J_4 + \frac{1}{2}M_2 M_p J_4 + \frac{1}{6}M_3 M_p J_5 \right). \tag{5.7}$$

On the other hand,

$$I_2 \leq \frac{M_2 s}{2} c_p \int_{\mathbb{R}^n \setminus B(0,r(x)/s)} \mathrm{e}^{-|v|^2/2}|v|^3\,\mathrm{d}v.$$

We have

$$\begin{aligned}
\|I_2\|_\rho^2 &\leq \left(\frac{M_2 s c_p}{2}\right)^2 \int_X \mathrm{d}\rho_X(x) \left(\int_{\mathbb{R}^n \setminus B(0,r(x)/s)} \mathrm{e}^{-|v|^2/2} |v|^3 \,\mathrm{d}v\right)^2 \\
&= \left(\frac{M_2 s c_p}{2}\right)^2 \int_0^{Diam(X)/2} \psi'(r)\,\mathrm{d}r \left(\frac{2\sqrt{\pi^n}}{\Gamma(n/2)} \int_{r/s}^{+\infty} t^{n+2} \mathrm{e}^{-t^2/2}\,\mathrm{d}t\right)^2 \\
&\leq \left(\frac{M_2 s c_p \sqrt{\pi^n}}{\Gamma(n/2)}\right)^2 M_{\psi'} s \int_0^{Diam(X)/2s} \mathrm{d}\xi \left(\int_\xi^{+\infty} t^{n+2} \mathrm{e}^{-t^2/2}\,\mathrm{d}t\right)^2,
\end{aligned}$$

where $\xi = r/s$, and we emphasize that the notation ξ is different from the one in the proof on Theorem 3. Also, u, x, y, r, and θ are temporarily employed in the following inequalities as integral variables only.

$$\begin{aligned}
& \int_0^{Diam(X)/2s} \mathrm{d}\xi \left(\int_\xi^{+\infty} t^{n+2} \mathrm{e}^{-t^2/2}\,\mathrm{d}t\right)^2 \\
\leq\ & \int_0^{+\infty} \mathrm{d}\xi \int_\xi^{+\infty} \int_\xi^{+\infty} x^{n+2} y^{n+2} \mathrm{e}^{-(x^2+y^2)/2}\,\mathrm{d}x\,\mathrm{d}y \\
\leq\ & \int_0^{+\infty} \mathrm{d}\xi \int_\xi^{+\infty} \mathrm{d}r \int_0^{\pi/2} r^{2(n+2)+1} \mathrm{e}^{-r^2/2} \cos^{n+2}\theta \sin^{n+2}\theta\ \mathrm{d}\theta \\
=\ & 2^{n+1} B\left(\frac{n+3}{2}, \frac{n+3}{2}\right) \int_0^{+\infty} \mathrm{d}\xi \int_\xi^{+\infty} \left(\frac{r^2}{2}\right)^{n+2} \mathrm{e}^{-r^2/2}\,\mathrm{d}\left(\frac{r^2}{2}\right).
\end{aligned}$$

Where $B(p,q) := 2\int_0^{\pi/2} \sin^{2p-1}\theta \cos^{2q-1}\theta\,\mathrm{d}\theta$ is the Euler-Beta function for any $p, q > 0$, and $B(p,q) = \Gamma(p)\Gamma(q)/\Gamma(p+q)$. So by putting $u = (r^2 - \xi^2)/2$,

$$\begin{aligned}
& \int_0^{+\infty} \mathrm{d}\xi \int_\xi^{+\infty} \left(\frac{r^2}{2}\right)^{n+2} \mathrm{e}^{-r^2/2}\,\mathrm{d}\left(\frac{r^2}{2}\right) \\
=\ & \int_0^{+\infty} \mathrm{d}\xi \int_0^{+\infty} \left(u + \frac{\xi^2}{2}\right)^{n+2} \mathrm{e}^{-u-\frac{\xi^2}{2}}\,\mathrm{d}u = \sum_{i=0}^{n+2} \frac{(n+2)!}{i! 2^i} \int_0^{+\infty} \mathrm{e}^{-\xi^2/2} \xi^{2i}\,\mathrm{d}\xi \\
=\ & \sum_{i=0}^{n+2} \frac{(n+2)!}{i! 2^i} 2^{(2i-1)/2} \Gamma\left(i + \frac{1}{2}\right) \leq (n+2)! \sum_{i=0}^{n+2} \frac{1}{\sqrt{2}} \leq (n+3)!\,.
\end{aligned}$$

Then we obtain

$$\|I_2\|_\rho^2 \leq \left(\frac{M_2 s c_p \sqrt{\pi^n}}{\Gamma(n/2)}\right)^2 M_{\psi'} s 2^{n+1} \frac{\Gamma((n+3)/2)^2 (n+3)!}{\Gamma(n+3)},$$

hence

$$\|I_2\|_\rho \le s^{3/2} M_2 c_p \sqrt{\pi^n M_{\psi'}} 2^{(n+1)/2} \Gamma\left(\frac{n+3}{2}\right) \sqrt{n+3} \Big/ \Gamma\left(\frac{n}{2}\right),$$

which, combined with (5.7), implies our result since $s \le 1$. □

We also need the following technical lemma.

Lemma 13. *For any $0 \le \tau < 1$ and $q \ge 2$,*

$$\sum_{i=2}^{q} \left((i-1)^{-\tau} \Big/ \sum_{j=i}^{q} j^{-\tau} \right) \le 6 \log \frac{q+1}{1-\tau}. \tag{5.8}$$

Proof. Denote the left hand side of (5.8) by Q, then

$$\begin{aligned} Q &\le 3 \sum_{i=1}^{q-1} \frac{i^{-\tau}}{\sum_{j=i}^{q} j^{-\tau}} \le 3 \sum_{i=1}^{q-1} \frac{i^{-\tau}(1-\tau)}{(q+1)^{1-\tau} - i^{1-\tau}} \\ &= \frac{3(1-\tau)}{q+1} \sum_{i=1}^{q-1} \frac{\left(\frac{i}{q+1}\right)^{-\tau}}{1 - \left(\frac{i}{q+1}\right)^{1-\tau}}. \end{aligned}$$

Let $t_i = \left(\frac{i}{q+1}\right)^{1-\tau}$ for $i = 1, \cdots, q$. For any $i \ge 1$, $(i+1)^{1-\tau} - i^{1-\tau} = (1-\tau)(i+\theta_i)^{-\tau}$, where $0 < \theta_i < 1$. Since $\left(\frac{i}{q+1}\right)^{-\tau} \ge 1$ for any $i = 1, \cdots, q$, we have

$$t_{i+1} - t_i = \frac{(1-\tau)(i+\theta_i)^{-\tau}}{(q+1)^{1-\tau}} \ge \frac{(1-\tau)(i+1)^{-\tau}}{(q+1)^{1-\tau}},$$

then

$$\frac{1}{q+1} \le \frac{(t_{i+1} - t_i)(1-\tau)^{-1}(i+1)^{\tau}}{(q+1)^{\tau}},$$

which implies

$$\begin{aligned} Q &\le 3 \sum_{i=1}^{q-1} \frac{\left(1+\frac{1}{i}\right)^{\tau} (t_{i+1} - t_i)}{1 - t_i} \le 6 \sum_{i=1}^{q-1} \frac{t_{i+1} - t_i}{1 - t_i} \\ &\le 6 \int_0^{\left(\frac{q}{q+1}\right)^{1-\tau}} \frac{\mathrm{d}x}{1-x} = 6 \log \left(\frac{(q+1)^{1-\tau}}{(q+1)^{1-\tau} - q^{1-\tau}} \right) \\ &\le 6 \log \left(\frac{(q+1)^{1-\tau}}{(1-\tau)(q+1)^{-\tau}} \right) = 6 \log \frac{q+1}{1-\tau}. \end{aligned}$$

The proof is thus completed. □

The following Lemma, also employed in [14], follows directly from the spectral decomposition, and the fact that $x\prod_{i=1}^{q}(1-\alpha_i x) \leq (\mathrm{e}\sum_{i=1}^{q}\alpha_i)^{-1}$ for any $0 \leq x \leq \min_{1\leq i\leq 1}\frac{1}{\alpha_i}$. We thus omit the proof.

Lemma 14. *Let $L \in \mathfrak{L}(H)$ be positive for some Hilbert space H. Suppose we have non-negative numbers $\alpha_1, \cdots, \alpha_q$, s.t. $\|L\| \cdot \max_{1\leq i\leq q}\alpha_i \leq 1$. Then*

$$\left\| \left(\prod_{i=1}^{q}(1-\alpha_i L) \right) L \right\| \leq \left(\mathrm{e}\sum_{i=1}^{q}\alpha_i \right)^{-1}.$$

□

Since L_K, Γ_s are positive on $\left(L^2_{\rho_X}\right)^n$, so is $\Gamma_s^{1/2}L_K\Gamma_s^{1/2}$. As was proved in (3.4), $\|\Gamma_s\|_\rho \leq c_p J_2$. On the other hand, for any $\vec{g}$ in $\left(L^2_{\rho_X}\right)^n$,

$$\begin{aligned} \|L_K\vec{g}\|_\rho^2 &= \int_X \mathrm{d}\rho_X(u)\left|\int_X \vec{g}(x)K(x,u)\,\mathrm{d}\rho_X(x)\right|^2 \\ &\leq \kappa^4\int_X \mathrm{d}\rho_X(u)\int_X |\vec{g}(x)|^2\,\mathrm{d}\rho_X(x) = \kappa^4\|\vec{g}\|_\rho^2. \end{aligned}$$

So, $\|L_K\|_\rho \leq \kappa^2$. We see that if $\gamma_1 \leq \left(\kappa^2 c_p J_2\right)^{-1}$,

$$\|\Gamma_s^{1/2}L_K\Gamma_s^{1/2}\|_\rho \max_{1\leq i\leq q}\gamma_i \leq 1, \tag{5.9}$$

for any $q \geq 1$. Base on the facts, we give the proof of Theorem 4.

Proof of Theorem 4. From the definition of iteration, one has

$$\vec{f}_{k+1} - \nabla f_\rho = (1-\gamma_k L_{K,s})\vec{f}_k - \nabla f_\rho + \gamma_k \vec{f}_{\rho,s}, \quad k = 1, 2, \cdots.$$

Since $\vec{f}_1 = 0$, direct computing shows

$$\begin{aligned} \vec{f}_{k+1} - \nabla f_\rho &= -\prod_{i=1}^{k}(1-\gamma_i L_{K,s})\nabla f_\rho \\ &\quad + \sum_{i=1}^{k}\gamma_i \prod_{p=i+1}^{k}(1-\gamma_p L_{K,s})(\vec{f}_{\rho,s} - L_{K,s}\nabla f_\rho) \\ &=: -H_1 + H_2. \end{aligned}$$

So,

$$H_1 = \Gamma_s^{-\frac{1}{2}} \left(\prod_{i=1}^{k} (1 - \gamma_i \Gamma_s^{\frac{1}{2}} L_K \Gamma_s^{\frac{1}{2}}) \right) \Gamma_s^{\frac{1}{2}} L_K \Gamma_s^{\frac{1}{2}} \Gamma_s^{-\frac{1}{2}} (L_K^{-1} \nabla f_\rho).$$

By (5.9) and Theorem 3, we get

$$\begin{aligned} \|H_1\|_\rho \leq \frac{1}{s^\alpha w} \cdot \frac{1}{\mathrm{e} \sum_{i=1}^{k} \gamma_i} \|L_K^{-1} \nabla f_\rho\|_\rho &\leq \frac{\|L_K^{-1} \nabla f_\rho\|_\rho (1-\tau)}{\mathrm{e} w s^\alpha \gamma_1 \left((k+1)^{1-\tau} - 1 \right)} \\ &\leq \frac{\|L_K^{-1} \nabla f_\rho\|_\rho (1-\tau)}{\mathrm{e} w s^\alpha \gamma_1 (1 - 2^{\tau-1})(k+1)^{1-\tau}}. \end{aligned}$$

On the other hand,

$$\begin{aligned} H_2 \;=\; & \Gamma_s^{-\frac{1}{2}} \sum_{i=1}^{k-1} \gamma_i \left(\prod_{q=i+1}^{k} (1 - \gamma_q \Gamma_s^{\frac{1}{2}} L_K \Gamma_s^{\frac{1}{2}}) \right) \Gamma_s^{\frac{1}{2}} L_K \Gamma_s^{\frac{1}{2}} \Gamma_s^{-\frac{1}{2}} (\vec{\zeta} - \Gamma_s \nabla f_\rho) \\ & + \gamma_k (\vec{f}_{\rho,s} - L_{K,s} \nabla f_\rho). \end{aligned}$$

So we have by Lemma 12 and Lemma 13

$$\begin{aligned} \|H_2\|_\rho \;&\leq\; \frac{1}{s^\alpha w} \sum_{i=1}^{k-1} i^{-\tau} \left(\mathrm{e} \sum_{j=i+1}^{k} j^{-\tau} \right)^{-1} C_5 s^{3/2} + \gamma_1 k^{-\tau} \kappa^2 C_5 s^{3/2} \\ &\leq\; \frac{6 C_5}{\mathrm{e} w} s^{\frac{3}{2} - \alpha} \log \frac{k+1}{1-\tau} + C_5 \kappa^2 \gamma_1 s^{3/2}, \end{aligned}$$

which finishes the proof. □

6. Proofs of the Main Results

Proof of Theorem 1. $m > (1-\tau)^{\frac{4\alpha - 22 - 8n}{2n+3}}$ implies

$$\left((1-\tau) m^{\left(n+\frac{3}{2}\right)/(4n+11-2\alpha)} \right)^{1/(1-\tau)} > 1,$$

and thus $k^* \geq 1$, So we have

$$\left((1-\tau) m^{\left(n+\frac{3}{2}\right)/(4n+11-2\alpha)} \right)^{1/(1-\tau)} \leq k^* + 1 \leq 2 \left((1-\tau) m^{\left(n+\frac{3}{2}\right)/(4n+11-2\alpha)} \right)^{1/(1-\tau)},$$

that is

$$m^{\left(n+\frac{3}{2}\right)/(4n+11-2\alpha)} \leq \frac{(k^*+1)^{1-\tau}}{1-\tau} \leq 2^{1-\tau} m^{\left(n+\frac{3}{2}\right)/(4n+11-2\alpha)}.$$

Then, we have by Lemma 2, Theorem 4, and inequality (3.2), with confidence at least $1-2\delta$ for any $\delta \in (0,1/2)$, that

$$\begin{aligned}
& \|\vec{f}_{k^*+1}^{\mathbf{z}} - \nabla f_\rho\|_\rho \le \kappa \|\vec{f}_{k^*+1}^{\mathbf{z}} - \vec{f}_{k^*+1}\|_{\mathcal{H}_K^n} + \|\vec{f}_{k^*+1} - \nabla f_\rho\|_\rho \\
\le & \ \frac{C_3\kappa(k^*+1)^{2-2\tau}}{s\sqrt{m}(1-\tau)^2}\log\frac{2}{\delta} + \frac{\|L_K^{-1}\nabla f_\rho\|_\rho(1-\tau)(1+c_pJ_1)\kappa^2}{\mathrm{e}ws^{\alpha+n}(1-2^{\tau-1})(k^*+1)^{1-\tau}} \\
& + \frac{C_5 s^{n+\frac{3}{2}}}{1+c_pJ_2} + \frac{6C_5}{\mathrm{e}w}s^{\frac{3}{2}-\alpha}\log\frac{k^*+1}{1-\tau},
\end{aligned}$$

so

$$\begin{aligned}
& \|\vec{f}_{k^*+1}^{\mathbf{z}} - \nabla f_\rho\|_\rho \le \frac{2^{2-2\tau}C_3\kappa}{s_0}m^{(-\frac{3}{2}+\alpha)/(4n+11-2\alpha)}\log\frac{2}{\delta} \\
& + \frac{\|L_K^{-1}\nabla f_\rho\|_\rho(1+c_pJ_1)\kappa^2}{\mathrm{e}ws_0^{\alpha+n}(1-2^{\tau-1})}m^{(-\frac{3}{2}+\alpha)/(4n+11-2\alpha)} + \frac{C_5 s_0^{n+\frac{3}{2}}}{1-c_pJ_2}m^{-(n+\frac{3}{2})/(4n+11-2\alpha)} \\
& + \frac{6C_5}{\mathrm{e}w}s_0^{\frac{3}{2}-\alpha}m^{(-\frac{3}{2}+\alpha)/(4n+11-2\alpha)}\left(\log 2 + \frac{\log m}{4(1-\tau)}\right) \\
& \le C_1 m^{(-\frac{3}{2}+\alpha)/(4n+11-2\alpha)}\left(1+\frac{\log m}{4(1-\tau)}\right)\log\frac{2}{\delta},
\end{aligned}$$

where

$$C_1 = \frac{2^{2-2\tau}C_3\kappa}{s_0} + \frac{\|L_K^{-1}\nabla f_\rho\|_\rho\kappa^2(1+c_pJ_2)}{\mathrm{e}ws_0^{n+\alpha}(1-2^{\tau-1})} + \frac{C_5 s_0^{n+\frac{3}{2}}}{1+c_pJ_2} + \frac{6C_5}{\mathrm{e}w}s_0^{\frac{3}{2}-\alpha}.$$

The proof of Theorem 1 is completed by replacing δ by $\delta/2$. □

Proof of Theorem 2. $m > (1-\tau)^{\frac{2\alpha-11}{3/2}}$ implies

$$\left((1-\tau)m^{\frac{3/2}{11-2\alpha}}\right)^{1/(1-\tau)} > 1,$$

and thus $k^* \ge 1$. So we have

$$\left((1-\tau)m^{\frac{3/2}{11-2\alpha}}\right)^{1/(1-\tau)} \le k^*+1 \le 2\left((1-\tau)m^{\frac{3/2}{11-2\alpha}}\right)^{1/(1-\tau)},$$

which is equivalent to

$$m^{\frac{3/2}{11-2\alpha}} \le \frac{(k^*+1)^{1-\tau}}{1-\tau} \le 2^{1-\tau}m^{\frac{3/2}{11-2\alpha}}. \tag{6.1}$$

By Theorem 4 and (4.17), for any $\delta \in (0, 1/3)$, we have with confidence $1-\delta$,

$$\begin{aligned}
&\|\vec{f}_{k^*+1} - \nabla f_\rho\|_\rho \\
\leq\ & \frac{3^\alpha \|L_K^{-1}\nabla f_\rho\|_\rho}{\mathrm{e}w\gamma_1(1-2^{\tau-1})C_4^\alpha \delta^{\alpha/n}} m^{\frac{-\frac{3}{2}+\alpha}{11-2\alpha}} \left(2\sqrt{\log m}+1\right)^\alpha \left(\log\frac{1}{\delta}\right)^{\frac{\alpha}{2}} \\
&+C_5\gamma_1\kappa^2 m^{\frac{-3/2}{11-2\alpha}+\frac{3}{n}} + \frac{6C_5}{\mathrm{e}w} m^{\frac{\alpha-\frac{3}{2}}{11-2\alpha}+\frac{2}{n}(\frac{3}{2}-\alpha)} \log\frac{k^*+1}{1-\tau},
\end{aligned}$$

then

$$\begin{aligned}
\|\vec{f}_{k^*+1} - \nabla f_\rho\|_\rho \leq & \left(\frac{3^\alpha \|L_K^{-1}\nabla f_\rho\|_\rho}{\mathrm{e}w\gamma_1(1-2^{\tau-1})C_4^\alpha} + C_5\gamma_1\kappa^2 + \frac{6C_5}{\mathrm{e}w}\right) \\
&\cdot m^{(\frac{3}{2}-\alpha)(\frac{-1}{11-2\alpha}+\frac{2}{n})} \left(2\sqrt{\frac{\log m}{1-\tau}}+1\right)^2 \delta^{-\frac{\alpha}{n}} \left(\log\frac{1}{\delta}\right)^{\frac{\alpha}{2}}, \qquad (6.2)
\end{aligned}$$

where we used

$$\begin{aligned}
\log\frac{k^*+1}{1-\tau} &\leq \frac{1}{1-\tau}\left(\log 2^{1-\tau} + \log m^{\frac{3/2}{11-2\alpha}}\right) \\
&\leq 1 + \frac{3}{16(1-\tau)}\log m \leq \left(1+2\sqrt{\frac{\log m}{1-\tau}}\right)^2.
\end{aligned}$$

By Lemma 3 and (3.2), we have with confidence $1-3\delta$,

$$\begin{aligned}
&\|\vec{f}^{\mathbf{z}}_{k^*+1} - \vec{f}_{k^*+1}\|_\rho \leq \kappa\|\vec{f}^{\mathbf{z}}_{k^*+1} - \vec{f}_{k^*+1}\|_{\mathcal{H}_K^n} \\
\leq\ & \frac{364M\left(2\sqrt{\log m}+1\right)}{\sqrt{n}C_4\delta^{1/n}} m^{\frac{2\alpha-9}{2(11-2\alpha)}+\frac{3}{11-2\alpha}} \left(\log\frac{2\mathrm{e}}{\delta}\right)^{\frac{5}{2}} \\
\leq\ & \frac{364M}{\sqrt{n}C_4} m^{(\frac{3}{2}-\alpha)(\frac{-1}{11-2\alpha}+\frac{2}{n})} \left(2\sqrt{\frac{\log m}{1-\tau}}+1\right)^2 \delta^{-\frac{1}{n}} \left(\log\frac{2\mathrm{e}}{\delta}\right)^{\frac{5}{2}}. \qquad (6.3)
\end{aligned}$$

Since (4.17) and (4.2) hold simultaneously with confidence $1-3\delta$, the proof is completed by combining (6.2) and (6.3) together, and replacing δ by $\delta/3$. The constant C_2 is defined as

$$C_2 = \frac{364M}{\sqrt{n}C_4} + \frac{3^\alpha \|L_K^{-1}\nabla f_\rho\|_\rho}{\mathrm{e}w\gamma_1(1-2^{\tau-1})C_4^\alpha} + C_5\gamma_1\kappa^2 + \frac{6C_5}{\mathrm{e}w}.$$

This proves Theorem 2. □

References

[1] R.A. Adams, Sobolev Spaces, Academic Press, New York, 1978.

[2] N. Aronszajn, Theory of Reproducing Kernels, Trans. Amer. Math. Soc. 68 (1950) 337-404.

[3] F. Cucker, S. Smale, On the Mathematical Foundations of Learning, Bull. Amer. Math. Soc. 39 (2001) 1-49.

[4] F. Cucker, D.X. Zhou, Learning Theory: an Approximation Theory Viewpoint, Cambridge University Press, 2007.

[5] X. Dong, D.X. Zhou, Learning Gradients by a Gradient Descent Algorithm, J. Math. Anal. Appl. 341 (2008) 1018-1027.

[6] N. Dunford, J.T. Schwartz, Linear Operators, Part II, Wiley, New York, 1988.

[7] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, E.S. Lander, Molecular Classification of Cancer: Class discovery and Class Prediction by Gene Expression Monitoring, Science 286 (1999) 531-537.

[8] S. Mukherjee, Q. Wu, Estimation of Gradients and Coordinate Covariation in Classification, J. Mach. Learn. Res. 7 (2006) 2481-2514.

[9] S. Mukherjee, Q. Wu, D.X. Zhou, Learning Gradients and Feature Selection on Manifolds, Bernoulli, to appear.

[10] S. Mukherjee, D.X. Zhou, Learning Coordinate Covariances via Gradients, J. Mach. Learn. Res. 7 (2006) 519-549.

[11] I. Pinelis, Optimum Bounds for the Distributions of Martingales in Banach Spaces, Ann. Probab. 22 (1994) 1679-1706.

[12] S. Smale, D.X. Zhou, Shannon sampling II: Connection to learning theory, Appl. Comput. Harmon. Anal. 19 (2005) 285-302.

[13] S. Smale, D.X. Zhou, Learning theory estimates via integral operators and their approximations, Constr. Approx. 26 (2007) 153-172.

[14] Y. Yao, L. Rosasco, A. Caponnetto, On Early Stopping in Gradient Descent Learning, Constr. Approx. 26 (2007) 289-315.

[15] Y. Ying, D.X. Zhou, Online Regularized Classification Algorithms, IEEE Trans. Inform. Theory 52 (2006) 4775-4788.