Chin. Ann. Math. Ser. B
39(4), 2018, 705–720
DOI: 10.1007/s11401-018-0091-7

Chinese Annals of Mathematics, Series B

Error Analysis on Hérmite Learning with Gradient Data*

Baohuai SHENG[1] Jianli WANG[1] Daohong XIANG[2]

Abstract This paper deals with Hérmite learning which aims at obtaining the target function from the samples of function values and the gradient values. Error analysis is conducted for these algorithms by means of approaches from convex analysis in the framework of multi-task vector learning and the improved learning rates are derived.

Keywords Hérmite learning, Gradient learning, Learning rate, Convex analysis, Multitask learning, Differentiable reproducing kernel Hilbert space
2000 MR Subject Classification 41A25, 46E22, 68Q32, 68T05, 90C25

1 Introduction

We considered Hérmite learning with gradient vectors (see [1–4]) in this paper. It can produce a smoothness function when the function value samples and gradient data is provided.

Let X be a compact subset of the Euclidean space $\mathbb{R}^d$ on which the learning or function approximation is considered. For each $x = (x^1, x^2, \cdots, x^d)^{\mathrm{T}} \in X$, the gradient of function $f : X \to R$ at x is denoted by the vector

$$\nabla_x f(x) = \Big(\frac{\partial f(x)}{\partial x^1}, \cdots, \frac{\partial f(x)}{\partial x^d}\Big)^{\mathrm{T}}$$

if the partial derivative for each variable exists. It is known (see [2]) that the essence of Hérmite learning is to obtain the regression function $f_\rho(x) = \int_R y \mathrm{d}\rho(y|x)$ from samples $z = \{(x_i, y_i)\}_{i=1}^m$ with $y_i = (y_i^0, \widetilde{y_i})$, where $\{(x_i, y_i^0)\}_{i=1}^m$ are drawn independently according to $\rho(x, y) = \rho(y|x)\rho_X(x)$ and $\widetilde{y_i} \approx \nabla f(x_i)$.

Let $K : X \times X \to R$ be a Mercer kernel which is continuous, symmetric and positive semi-definite, which means that the matrix $(K(x_i, x_j))_{i,j=1}^l$ is positive semi-definite for any finite set of points $\{x_1, \cdots, x_l\} \subset X$. The associated reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$ is defined (see [5–7]) as the completion of the linear span of the set of the function $\{K_x = K(x, \cdot) : x \in X\}$ with the inner product given by

$$\langle K_x, K_y \rangle_K = K(x, y).$$

Let $L^2(\mu)$ be the class of all square integrable functions with respect to the mesure μ with the norm

$$\|f\|_{L_2(\mu)} = \Big(\int_X |f(x)|^2 \, \mathrm{d}\mu\Big)^{\frac{1}{2}} < +\infty.$$

Manuscript received February 9, 2015. Revised May 28, 2016.
[1]Department of Mathematics, Shaoxing Univeristy, Shaoxing 312000, Zhejiang, China.
E-mail: bhsheng@usx.edu.cn wjl@usx.edu.cn
[2]Department of Mathematics, Zhejiang Normal University, Jinhua 321004, Zhejiang, China.
E-mail: daohongxiang@zjnu.cn
*This work was supported by the National Natural Science Foundation of China (No. 11471292).

Define the integral operator $\mathcal{L}_K$ as

$$\mathcal{L}_K(f,x)=\int_X f(u)K(x,u)\,\mathrm{d}\mu,\quad f\in L^2(\mu).$$

Since K is a positive semi-definite, $\mathcal{L}_K$ is a compact positive operator. Let λ_k be the k-th positive eigenvalue of $\mathcal{L}_K(f)$ and $\phi_k(x)$ be the corresponding continuous orthonormal eigenfunction. Then, by Mercer's theorem, for all $x,y\in X$, there holds

$$K(x,y)=\sum_{k=0}^{+\infty}\lambda_k\ \phi_k(x)\ \phi_k(y),\quad x,y\in X, \tag{1.1}$$

where the convergence is absolute (for each $x,\ y\in X$) and uniform on $X\times X$. Then we know (see [8–9])

$$\mathcal{H}_K=\mathcal{L}_K^{\frac{1}{2}}(L_2(\mu))=\Big\{f(x)=\sum_{k=0}^{\infty}a_k(f)\phi_k(x):\ \sum_{k=0}^{\infty}\frac{|a_k(f)|^2}{\lambda_k}<+\infty\Big\}$$

with inner product

$$\langle f,\ g\rangle_{\mathcal{H}_K}=\sum_{k=0}^{\infty}\frac{a_k(f)\ a_k(g)}{\lambda_k}$$

and $a_k(f)=\int_X f(y)\ \phi_k(y)\,\mathrm{d}\mu(y)$.

When X is a compact set, there is a constant $k>0$ such that

$$\|f\|_{\infty,X}\le k\|f\|_{\mathcal{H}_K},\quad f\in\mathcal{H}_K. \tag{1.2}$$

Let $M>0$ be a given positive real number and $Y=[-M,M]^{d+1}$. We denote by $y\in Y$ as $y=(y^0,\widetilde{y})$ with $\widetilde{y}=(y^1,y^2,\cdots,y^d)^{\mathrm{T}}$ and

$$\overrightarrow{f_\rho}(x)=(f_\rho(x),\widetilde{f}_\rho(x)),$$

where

$$f_\rho(x)=\int_{[-M,\ M]}y^0\,\mathrm{d}\rho(y|x)$$

and

$$\begin{aligned}\widetilde{f}_\rho(x)&=\int_{[-M,\ M]^d}\widetilde{y}\,\mathrm{d}\rho(y|x)\\&=\Big(\int_{[-M,\ M]^d}y^1\,\mathrm{d}\rho(y|x),\cdots,\int_{[-M,\ M]^d}y^d\,\mathrm{d}\rho(y|x)\Big)^{\mathrm{T}}.\end{aligned}$$

Let $y_i^0\in[-M,\ M]$ and $\widetilde{y}_i=(y_i^1,\cdots,y_i^d)^{\mathrm{T}}\in[-M,\ M]^d$. Then, the Hérmite learning algorithm corresponding to the samples z is (see [1])

$$f_{z,\lambda}:=\arg\min_{f\in\mathcal{H}_K}\frac{1}{m}\sum_{i=1}^{m}((y_i^0-f(x_i))^2+\|\widetilde{y}_i-\nabla f(x_i)\|_{\mathbb{R}^d}^2)+\lambda\,\|f\|_{\mathcal{H}_K}^2, \tag{1.3}$$

where $\lambda>0$ is the regularization parameter, $\|\cdot\|_{\mathbb{R}^d}$ is the usual norm of the Euclidean space $\mathbb{R}^d$.

The representer theorem for scheme (1.3) was provided in [1], which showed that the minimization over the possibly infinite dimensional space $\mathcal{H}_K$ can be achieved in a finite dimensional subspace generated by $\{K_{x_i}(\cdot)\}$ and their partial derivatives. The explicit solution to (1.3) and an upper bound for the learning rate with the integral operator approach were given in [2].

We notice that the problem of learning multiple tasks with kernel methods has been a developing topic (see [10–14]). The representer theorem for these frameworks have been given, but the results of convergence quantitative analysis are relatively few, and very effective methods for bounding the convergence rates does not appear. An aim of the present paper is to show that model (1.3) is among the group of the multiple tasks. For this purpose, we rewrite model (1.3) from the view of vector valued functions.

Let $C^{(1)}$ be the class of all the real functions $f(x)$ defined on X with continuous partial derivatives and $C^{(2)}(X)$ denote the class of all the real functions $f(x)$ defined on X with continuous partial derivatives $\frac{\partial^2 f}{\partial x^\alpha \partial y^\beta}$ for $\alpha = 1, 2, \cdots, d$. By $K \in C^{(2)}(X \times X)$, we mean all the partial derivatives $\frac{\partial^2 \partial^2 K(x,y)}{\partial (x^\alpha)^2 \partial (x^\beta)^2}$ that are continuous on $X \times X$. By [15–17] we know that if the kernels $K(x, y)$ have the form of (1.1) and $K(x, y) \in C^{(2)}(X \times X)$, then, $\mathcal{H}_K$ can be embedded into both $C^{(1)}(X)$ and $C^{(2)}(X)$, and for any given $x \in X$ and all $\alpha = 1, 2, \cdots, d$, the following relations hold:

$$\frac{\partial K_x(\cdot)}{\partial x^\alpha} \in \mathcal{H}_K, \quad \frac{\partial^2 K_x(\cdot)}{\partial (x^\alpha)^2} \in \mathcal{H}_K, \tag{1.4}$$

$$\frac{\partial f(x)}{\partial x^\alpha} = \Big\langle f, \ \frac{\partial}{\partial x^\alpha} K_x(\cdot) \Big\rangle_{\mathcal{H}_K}, \quad f \in \mathcal{H}_K \tag{1.5}$$

and

$$\Big| \frac{\partial f(x)}{\partial x^\alpha} \Big| \le k_1 \|f\|_{\mathcal{H}_K}, \qquad \forall x \in X, \ \forall \alpha = 0, 1, 2, \cdots, d, \tag{1.6}$$

where $\frac{\partial f}{\partial x^0} = f(x), \ k_1 = \sup\limits_{\substack{x,y \in X \\ 0 \le \alpha, \beta \le d}} \sqrt{\Big| \frac{\partial^2 K(x,y)}{\partial x^\alpha \partial y^\beta} \Big|}$.

Define

$$\mathcal{H}_K^\alpha = \Big\{ f_\alpha(x) = \frac{\partial f(x)}{\partial x^\alpha} : \ f(x) \in \mathcal{H}_K \Big\}, \quad \alpha = 1, 2, \cdots, d$$

and

$$\begin{aligned} \overrightarrow{\mathcal{H}_K} &= \mathcal{H}_K \times \mathcal{H}_K^1 \times \cdots \times \mathcal{H}_K^d \\ &= \{ \overrightarrow{f} = (f, f_1, \cdots, f_d)^{\mathrm{T}} : \ f \in \mathcal{H}_K, \ f_\alpha \in \mathcal{H}_K^\alpha, \ \alpha = 1, 2, \cdots, d \}. \end{aligned}$$

Then, (1.3) can be rewritten as

$$\overrightarrow{f_{z,\lambda}} := \arg \min_{\overrightarrow{f} = (f, f_1, \cdots, f_d)^{\mathrm{T}} \in \overrightarrow{\mathcal{H}_K}} \mathcal{E}_z(\overrightarrow{f}) + \lambda \ \|f\|_{\mathcal{H}_K}^2, \tag{1.7}$$

where

$$\begin{aligned} &\mathcal{E}_z(\overrightarrow{f}) = \frac{1}{m} \sum_{i=1}^m \|\overrightarrow{y_i} - \overrightarrow{f}(x_i)\|_{\mathbb{R}^{d+1}}^2, \\ &\overrightarrow{y}_i = (y_i^0, \ y_i^1, \ \cdots, \ y_i^d)^{\mathrm{T}}, \qquad i = 1, 2, \cdots, m \end{aligned}$$

and

$$\overrightarrow{f_{z,\lambda}} = (f^{(z,\lambda)}, f_1^{(z,\lambda)}, \cdots, f_d^{(z,\lambda)})^{\mathrm{T}}.$$

Algorithm (1.7) is neither the same as the usual least square regression (see [18–21]), nor the current multitask learning model since it uses the penalty $\|f\|^2_{\mathcal{H}_K}$ not $\|\overrightarrow{f}\|^2_{\mathcal{H}_K}$. However, it is really a concrete example of multitask learning models, the study method of its performance can provide useful reference for the research of other multitask models. This is the first motivation for writing this paper. On the other hand, such approaches may provide a way of thinking for dealing with other gradient learning models. The structure of model (1.7) is close to the usual least square learning models, this fact creates an opportunity of studying the gradient learning with existing methods, e.g., the convex analysis method (see [20–24]). This is the second motivation for writing this paper.

We form an improved convex method with the help of Gâteaux derivative and the optimality conditions of convex functions and use it to bound the convergence rates of model (1.7). To show the main results of the present paper, we restate the notion of covering number.

For a distance space S and a real number $\eta > 0$. The covering number $\mathcal{N}(S, \eta)$ is defined to be the minimal positive integer number l such that there exists l disks in S with radius η covering S.

We call a compact subset E of a distance space $(\mathcal{B}, \|\cdot\|_{\mathcal{B}})$ logarithmic complexity exponent $s \geq 0$ if there is a constant $c_s > 0$ such that the closed ball of radius R centered at origin, i.e.,

$$\mathcal{B}_R = \{f \in E : \|f\|_{\mathcal{B}} \leq R\}$$

satisfies

$$\log \mathcal{N}(\mathcal{B}_R, \eta) \leq c_s \Big(\frac{R}{\eta}\Big)^s, \quad \forall \eta > 0. \tag{1.8}$$

We now give the following Theorem 1.1.

Theorem 1.1 *Let $K(x, y)$ be a Mercer-like kernel satisfying $K(x, y) \in C^{(2)}(X \times X)$, $\overrightarrow{f_{z,\lambda}}$ be the solution of* (1.7) *and $C = 512\, k_1^4\, (d+1)^3 M$. If $\mathcal{H}_k$ has logarithmic complexity exponent $s \geq 0$ in the uniform continuity norm $\|\cdot\|_{C(X)}$ and $\lambda \leq \frac{k_1^2\, d}{M^2} \times D(\overrightarrow{f_\rho}, \lambda)$, then, for any $0 < \delta < 1$ and $0 < \delta < \frac{4}{\mathrm{e}^{c_s}}$, with confidence $1 - \delta$, we have*

$$\|\overrightarrow{f_{z,\lambda}} - \overrightarrow{f_\rho}\|^2_{L^2(\rho_X)} \leq \frac{C\sqrt{D(\overrightarrow{f_\rho}, \lambda)}}{\lambda^2 \sqrt{m}} \times \Big(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt[2+s]{m}}\Big)\Big(\log \frac{4}{\delta}\Big)^2 + D(\overrightarrow{f_\rho}, \lambda), \tag{1.9}$$

where

$$D(\overrightarrow{f_\rho}, \lambda) = \inf_{\overrightarrow{h} \in \overrightarrow{\mathcal{H}_K}} (\mathcal{E}_\rho(\overrightarrow{h}) - \mathcal{E}_\rho(\overrightarrow{f_\rho}) + \lambda \|h\|^2_{\mathcal{H}_K})$$

and

$$\mathcal{E}_\rho(\overrightarrow{f}) = \int_Z \|\overrightarrow{y} - \overrightarrow{f}(x)\|^2_{\mathbb{R}^{d+1}}\, \mathrm{d}\rho.$$

We now give some comments on (1.9).

(1) Let $H^1_{\rho_X}$ be the Sobolev space consisting of all the functions $f \in L^2(\rho_X)$ with all partial derivatives belonging to $L^2(\rho_X)$, whose norm $\|f\|_{H^1_{\rho_X}}$ is induced by the inner product

$$\langle f,\, g\rangle_{H^1_{\rho_X}} = \int_X (f(x)\, g(x) + \nabla_x f(x) \nabla_x g(x))\, \mathrm{d}\rho_X.$$

If ρ is perfect (see [2]), i.e., $\widetilde{f_\rho}(x) = \nabla f_\rho(x)$, then, $\overrightarrow{f_\rho}(x) = (f_\rho(x), \nabla f_\rho(x))$ and

$$\|\overrightarrow{f_{z,\lambda}} - \overrightarrow{f_\rho}\|_{L^2(\rho_X)} = \|f_{z,\lambda} - f_\rho\|_{H^1_{\rho_X}}.$$

(2) By Lemma 2.2 afterward we have

$$D(\overrightarrow{f_\rho}, \lambda) = \inf_{\overrightarrow{h} \in \overrightarrow{\mathcal{H}_K}} (\|\overrightarrow{f_\rho} - \overrightarrow{h}\|^2_{L^2(\rho_X)} + \lambda \|h\|^2_{\mathcal{H}_K}).$$

As usual, we assume that there are given constants $c > 0$ and $0 < \beta < 4$ such that $D(\overrightarrow{f}_\rho, \lambda) \leq c\lambda^\beta$ if $\overrightarrow{\mathcal{H}_K}$ is density in $L^2(\rho_X)$. In this case, if $s = 0$, then we have by (1.9) that

$$\|f_{z,\lambda} - f_\rho\|^2_{H^1_{\rho,X}} = O\Big(\frac{1}{m\ \lambda^{2-\frac{\beta}{2}}} \times \Big(\log\frac{4}{\delta}\Big)^2 + \lambda^\beta\Big). \tag{1.10}$$

Further, if $\lambda = m^{-\theta}$ and $0 < \theta < \frac{2}{4-\beta}$, then, we have by (1.10) the following estimate

$$\|f_{z,\lambda} - f_\rho\|^2_{H^1_{\rho_X}} = O\Big(\frac{(\log\frac{4}{\delta})^2}{m^{1-(2-\frac{\beta}{2})\theta}} + m^{-\beta\theta}\Big). \tag{1.11}$$

(3) Define a new integral operator

$$L = L_K :\ H^1_{\rho_X} \to H^1_{\rho_X}$$

associated with K and ρ_X by

$$L_K(f,x) = \int_X (K(x,y)\ f(y) + \nabla_y K(x,y) \nabla_y f(y))\ \mathrm{d}\rho_X(y), \qquad x \in X,\ f \in H^1_{\rho_X}.$$

Let L^r_K be defined as

$$L^r_K\Big(\sum c_k\ \phi_k(x)\Big) = \sum c_k\ \lambda^r_k\ \phi_k(x)$$

for $r > 0$. Then, by taking $\lambda = 8(d+1)\ k^2\ \log(\frac{4}{\delta})\ m^{-\beta}$, with confidence $1-\delta$, we have (see [2])

$$\|f_{z,\lambda} - f_\rho\|_{H^1_{\rho_X}} \leq 8\ \log\Big(\frac{4}{\delta}\Big)\{M + (d+1)\ \|K\|^r_{C^2} \times \|L_K^{-r} f_\rho\|_{H^1_{\rho_X}}\}\ m^{-r\beta}, \tag{1.12}$$

where

$$\beta = \begin{cases} \dfrac{1}{2r+1}, & \text{if } \ r > \dfrac{1}{2}, \\ \dfrac{1}{2}, & \text{if } 0 < r \leq \dfrac{1}{2}. \end{cases}$$

It is easy to see that (1.11) sharps (1.12).

2 Proofs

To show Theorem 1.1, we need the notation of the Gâteaux derivative and some related lemmas.

Let $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ be a Hilbert space, $F(f) : \mathcal{H} \to R \cup \{\mp\infty\}$ be a real function. We say F is Gâteaux differentiable at $f \in \mathcal{H}$, if there is a $\xi \in \mathcal{H}$ such that for any $g \in \mathcal{H}$, there holds

$$\lim_{t\to 0} \frac{F(f+tg) - F(f)}{t} = \langle g, \xi\rangle_{\mathcal{H}} \tag{2.1}$$

and write $\nabla_f\ F(f) = \xi$ as the Gâteaux derivative of $F(f)$ at f.

Lemma 2.1 *Let $F(f): \mathcal{H} \to R \cup \{\mp\infty\}$ be a function defined on Hilbert space $\mathcal{H}$. Then, we have following results:*

(i) *If $F(f)$ is a convex function, then, $F(f)$ attains minimal value at f_0 if and only if $\nabla_f F(f_0) = 0$.*

(ii) *If $F(f): \mathcal{H} \to R \cup \{\mp\infty\}$ is a Gâteaux differentiable function, then, $F(f)$ is a convex on $\mathcal{H}$ if and only if for any $f, g \in \mathcal{H}$ we have*

$$F(g+f) - F(f) \geq \langle g, \nabla_f F(f)\rangle_{\mathcal{H}}. \tag{2.2}$$

Proof We have (i) from Proposition 17.4 of [25] and we have (ii) from Proposition 17.10 and Proposition 17.12 of [25].

Lemma 2.2 *$\overrightarrow{f_\rho}$ satisfies the relation*

$$\overrightarrow{f_\rho} = \arg \min_{\overrightarrow{f}=(f,f_1,\cdots,f_d)^{\mathrm{T}}} \mathcal{E}_\rho(\overrightarrow{f}) \tag{2.3}$$

and the equation

$$\mathcal{E}_\rho(\overrightarrow{f}) - \mathcal{E}_\rho(\overrightarrow{f_\rho}) = \int_Z \|\overrightarrow{f_\rho}(x) - \overrightarrow{f}(x)\|^2_{\mathbb{R}^{d+1}}\, \mathrm{d}\rho_X. \tag{2.4}$$

Proof Since equality

$$\|a+b\|^2_{\mathbb{R}^{d+1}} = \|a\|^2_{\mathbb{R}^{d+1}} + 2\langle a, b\rangle_{\mathbb{R}^{d+1}} + \|b\|^2_{\mathbb{R}^{d+1}}$$

holds for any $a, b \in \mathbb{R}^{d+1}$, we have

$$\begin{aligned}
\mathcal{E}_\rho(\overrightarrow{f}) &= \int_Z \|\overrightarrow{y} - \overrightarrow{f_\rho}\|^2_{\mathbb{R}^{d+1}}\, \mathrm{d}\rho - 2\int_Z \langle \overrightarrow{y} - \overrightarrow{f_\rho}(x),\ \overrightarrow{f}(x) - \overrightarrow{f_\rho}(x)\rangle_{\mathbb{R}^{d+1}}\, \mathrm{d}\rho \\
&\quad + \int_Z \|\overrightarrow{f}_\rho(x) - \overrightarrow{f}(x)\|^2_{\mathbb{R}^{d+1}} \mathrm{d}\rho \\
&= \mathcal{E}_\rho(\overrightarrow{f_\rho}) - 2\int_X \Big\langle \int_Y (\overrightarrow{y} - \overrightarrow{f_\rho}(x))\, \mathrm{d}\rho(y|x),\ \overrightarrow{f}(x) - \overrightarrow{f_\rho}(x)\Big\rangle_{\mathbb{R}^{d+1}}\, \mathrm{d}\rho_X \\
&\quad + \int_Z \|\overrightarrow{f_\rho}(x) - \overrightarrow{f}(x)\|^2_{\mathbb{R}^{d+1}}\ \mathrm{d}\rho_X \\
&= \mathcal{E}_\rho(\overrightarrow{f_\rho}) + \int_Z \|\overrightarrow{f_\rho}(x) - \overrightarrow{f}(x)\|^2_{\mathbb{R}^{d+1}}\, \mathrm{d}\rho_X.
\end{aligned} \tag{2.5}$$

Let $\overrightarrow{h_\lambda} = (h^{(\lambda)},\ h_1^{(\lambda)}, \cdots, h_d^{(\lambda)})^{\mathrm{T}}$ be defined as

$$\overrightarrow{h_\lambda} = \arg \min_{\overrightarrow{h}=(h,\ h_1,\ \cdots,\ h_d)^{\mathrm{T}} \in \overrightarrow{\mathcal{H}_K}} (\mathcal{E}_\rho(\overrightarrow{h}) + \lambda\, \|h\|^2_{\mathcal{H}_K}). \tag{2.6}$$

Then, we have the following Lemma 2.3.

Lemma 2.3 *Let $\overrightarrow{h_\lambda}$ be a solution of* (2.6)*. Then, we have*

$$\nabla_{\overrightarrow{f}}(\mathcal{E}_\rho(\overrightarrow{f}))(\cdot) = -2\int_Z \langle \nabla_x^* K_x(\cdot),\ \overrightarrow{y} - \overrightarrow{f}(x)\rangle_{\mathbb{R}^{d+1}}\, \mathrm{d}\rho \tag{2.7}$$

and

$$\nabla_{\overrightarrow{f}}(\mathcal{E}_z(\overrightarrow{f}))(\cdot) = -\frac{2}{m}\sum_{i=1}^m \langle \nabla_x^* K_{x_i}(\cdot),\ \overrightarrow{y}_i - \overrightarrow{f}(x_i)\rangle_{\mathbb{R}^{d+1}}, \tag{2.8}$$

where $\nabla_x^* K_x(\cdot) = \big(\frac{\partial K_x(\cdot)}{\partial x^0},\ \frac{\partial K_x(\cdot)}{\partial x^1},\ \cdots,\ \frac{\partial K_x(\cdot)}{\partial x^d}\big)^{\mathrm{T}}$ and $\frac{\partial K_x(\cdot)}{\partial x^0} = K_x(\cdot)$.

Proof By the equality

$$a^2 - b^2 = \langle a - b, \quad 2b\rangle_{\mathbb{R}^{d+1}} + \|a - b\|^2_{\mathbb{R}^{d+1}}, \quad a \in \mathbb{R}^{d+1},\ b \in \mathbb{R}^{d+1}. \tag{2.9}$$

we have

$$\begin{aligned}
&\lim_{t\to 0} \frac{\mathcal{E}_\rho(\overrightarrow{f} + t\overrightarrow{g}) - \mathcal{E}_\rho(\overrightarrow{f})}{t} \\
&= \lim_{t\to 0} \frac{1}{t}\Big(\int_Z \|\overrightarrow{y} - (\overrightarrow{f}(x) + t\overrightarrow{g}(x))\|^2_{\mathbb{R}^{d+1}}\, \mathrm{d}\rho - \int_Z \|\overrightarrow{y} - \overrightarrow{f}(x)\|^2_{\mathbb{R}^{d+1}}\, \mathrm{d}\rho\Big) \\
&= -2\int_Z \langle \overrightarrow{g}(x),\ \overrightarrow{y} - \overrightarrow{f}(x)\rangle_{\mathbb{R}^{d+1}}\, \mathrm{d}\rho.
\end{aligned}$$

Take $g_0(x) = g(x),\ f_0(x) = f(x)$. Then, by (1.5) we have

$$\begin{aligned}
\langle \overrightarrow{g}(x),\ \overrightarrow{y} - \overrightarrow{f}(x)\rangle_{\mathbb{R}^{d+1}} &= \sum_{k=0}^{d} g_k(x) \times (y^k - f_k(x)) \\
&= \sum_{k=0}^{d} \Big(g,\ \frac{\partial K_x(\cdot)}{\partial x^k}\Big)_{\mathcal{H}_K} \times (y^k - f_k(x)) \\
&= \Big(g, \sum_{k=0}^{d} \frac{\partial K_x(\cdot)}{\partial x^k}(y^k - f_k(x))\Big)_{\mathcal{H}_K} \\
&= (g,\ \langle \nabla_x^* K_x(\cdot),\ \overrightarrow{y} - \overrightarrow{f}(x)\rangle_{\mathbb{R}^{d+1}})_{\mathcal{H}_K}.
\end{aligned} \tag{2.10}$$

It follows

$$\lim_{t\to 0} \frac{\mathcal{E}_\rho(\overrightarrow{f} + t\overrightarrow{g}) - \mathcal{E}_\rho(\overrightarrow{f})}{t} = \Big(g,\ 2\int_Z \langle \nabla_x^* K_x(\cdot), \overrightarrow{y} - \overrightarrow{f}(x)\rangle_{\mathbb{R}^{d+1}} \mathrm{d}\rho\Big)_{\mathcal{H}_K}.$$

(2.7) thus holds. (2.8) can be proved in the same way.

Lemma 2.4 *Let $\overrightarrow{h_\lambda}$ be a solution of* (2.6) *and let $\overrightarrow{f_{z,\lambda}}$ be a solution of* (1.7)*. Then, we have*

$$\lambda h^{(\lambda)}(\cdot) = \int_Z \langle \nabla_x^* K_x(\cdot),\ \overrightarrow{y} - \overrightarrow{h_\lambda}(x)\rangle_{\mathbb{R}^{d+1}}\, \mathrm{d}\rho \tag{2.11}$$

and

$$\lambda f^{(z,\lambda)}(\cdot) = \frac{1}{m}\sum_{i=1}^{m} \langle \nabla_x^* K_{x_i}(\cdot),\ \overrightarrow{y}_i - \overrightarrow{f_{z,\lambda}}(x_i)\rangle_{\mathbb{R}^{d+1}}. \tag{2.12}$$

Proof Since $\overrightarrow{h}_\lambda$ is the solution of (2.6), we have

$$\begin{aligned}
&\nabla_{\overrightarrow{h}}(\mathcal{E}_\rho(\overrightarrow{h}) + \lambda\|h\|^2_{\mathcal{H}_K})|_{\overrightarrow{h}=\overrightarrow{h_\lambda}} \\
&= \nabla_{\overrightarrow{h}}\mathcal{E}_\rho(\overrightarrow{h})|_{\overrightarrow{h}=\overrightarrow{h_\lambda}} + \lambda\ \nabla_{\overrightarrow{h}}(\|h\|^2_{\mathcal{H}_K})|_{\overrightarrow{h}=\overrightarrow{h_\lambda}} \\
&= -2\int_Z \langle \nabla_x^* K_x(\cdot),\ \overrightarrow{y} - \overrightarrow{h_\lambda}(x)\rangle_{\mathbb{R}^{d+1}}\, \mathrm{d}\rho(x,y) + 2\lambda\, h^{(\lambda)}(\cdot) = 0,
\end{aligned} \tag{2.13}$$

where we have used the fact that

$$\nabla_{\overrightarrow{h}}(\|h\|^2_{\mathcal{H}_K}) = \nabla_h(\|h\|^2_{\mathcal{H}_K}) = 2h.$$

(2.11) thus holds. (2.12) can be proved in the same way.

Lemma 2.5 (2.6) *has a unique solution* $\overrightarrow{h_\lambda}$ *and* (1.7) *has a unique solution* $\overrightarrow{f_{z,\lambda}}$. *They satisfy the following inequalities:*

$$\|h^{(\lambda)}\|_{\mathcal{H}_K} \le \sqrt{\frac{D(\overrightarrow{f_\rho},\lambda)}{\lambda}} \tag{2.14}$$

and

$$\|f^{(z,\lambda)}\|_{\mathcal{H}_K} \le \frac{M}{\sqrt{\lambda}}. \tag{2.15}$$

Proof Let $\overrightarrow{f} = (f, f_1, \cdots, f_d)^{\mathrm{T}} \in \overrightarrow{\mathcal{H}_K}$ and $\overrightarrow{g} = (g,\ g_1,\ \cdots,\ g_d)^{\mathrm{T}} \in \overrightarrow{\mathcal{H}_K}$. Then, by the equality (2.9) we have

$$\begin{aligned}\mathcal{E}_\rho(\overrightarrow{g}) - \mathcal{E}_\rho(\overrightarrow{f}) &= \int_Z (\|\overrightarrow{y} - \overrightarrow{g}(x)\|^2_{\mathbb{R}^{d+1}} - \|\overrightarrow{y} - \overrightarrow{f}(x)\|^2_{\mathbb{R}^{d+1}})\mathrm{d}\rho \\ &= -2\int_Z \langle \overrightarrow{g}(x) - \overrightarrow{f}(x), \overrightarrow{y} - \overrightarrow{f}(x)\rangle_{\mathbb{R}^{d+1}}\mathrm{d}\rho + \int_Z \|\overrightarrow{f}(x) - \overrightarrow{g}(x)\|^2_{\mathbb{R}^{d+1}}\mathrm{d}\rho \\ &\ge -2\int_Z \langle \overrightarrow{g}(x) - \overrightarrow{f}(x), \overrightarrow{y} - \overrightarrow{f}(x)\rangle_{\mathbb{R}^{d+1}}\mathrm{d}\rho.\end{aligned}$$

Since (2.10), we have

$$\langle \overrightarrow{g}(x) - \overrightarrow{f}(x), \overrightarrow{y} - \overrightarrow{f}(x)\rangle_{\mathbb{R}^{d+1}} = (g - f,\ \langle \nabla^*_x K_x(\cdot),\ \overrightarrow{y} - \overrightarrow{f}(x)\rangle_{\mathbb{R}^{d+1}})_{\mathcal{H}_K}.$$

Therefore, by (2.7) we have

$$\begin{aligned}\mathcal{E}_\rho(\overrightarrow{g}) - \mathcal{E}_\rho(\overrightarrow{f}) &\ge \Big(g - f,\ -2\int_Z \langle \nabla^*_x K_x(\cdot),\ \overrightarrow{y} - \overrightarrow{f}(x)\rangle_{\mathbb{R}^{d+1}}\mathrm{d}\rho\Big)_{\mathcal{H}_K} \\ &= (g - f,\ \nabla_{\overrightarrow{f}}(\mathcal{E}_\rho(\overrightarrow{f})))_{\mathcal{H}_K}.\end{aligned}$$

By (ii) of Lemma 2.1 we know $\mathcal{E}_\rho(\overrightarrow{f})$ is a convex function on $\overrightarrow{\mathcal{H}_K}$. Since $\|f\|^2_{\mathcal{H}_K}$ is a strictly convex function on $\overrightarrow{\mathcal{H}_K}$ and $\lambda > 0$, we know

$$\mathcal{E}_\rho(\overrightarrow{h}) + \lambda\ \|h\|^2_{\mathcal{H}_K}$$

is strictly convex on $\overrightarrow{\mathcal{H}_K}$ and the optimal solution $\overrightarrow{h_\lambda}$ is unique. By the same way we can show the uniqueness of $\overrightarrow{f_{z,\lambda}}$.

By the definition of $\overrightarrow{h_\lambda}$ we have

$$\lambda\|h^{(\lambda)}\|^2_{\mathcal{H}_K} \le \mathcal{E}_\rho(\overrightarrow{h_\lambda}) - \mathcal{E}_\rho(\overrightarrow{f_\rho}) + \lambda\|h_\lambda\|^2_{\mathcal{H}_K} = D(\overrightarrow{f_\rho},\ \lambda).$$

(2.14) then holds. By the definition of $\overrightarrow{f_{z,\lambda}}$ we have

$$\mathcal{E}_z(\overrightarrow{f_{z,\lambda}}) + \lambda\|f_{z,\lambda}\|^2_{\mathcal{H}_K} \le \mathcal{E}_z(\overrightarrow{0}) = \frac{1}{m}\sum_{i=1}^m \|\overrightarrow{y}_i\|^2_{\mathbb{R}^{d+1}} \le M^2.$$

(2.15) thus holds.

Lemma 2.6 *Let $\overrightarrow{h_\lambda}$ be defined as* (2.6) *and let $\overrightarrow{f_{z,\lambda}}$ be the solution of* (1.7)*. Then, there holds*

$$\|h^{(\lambda)} - f^{(z,\lambda)}\|_{\mathcal{H}_K} \le \frac{2}{\lambda} \Big\| \int_Z \langle \nabla_x^* K_x(\cdot),\ \overrightarrow{y} - \overrightarrow{h_\lambda}(x)\rangle_{\mathbb{R}^{d+1}}\, \mathrm{d}\rho - \frac{1}{m}\sum_{i=1}^m \langle \nabla_x^* K_{x_i}(\cdot),\ \overrightarrow{y}_i - \overrightarrow{h_\lambda}(x_i)\rangle_{\mathbb{R}^{d+1}} \Big\|_{\mathcal{H}_K}. \tag{2.16}$$

Proof By (2.9), we have

$$\begin{aligned}\mathcal{E}_z(\overrightarrow{f_{z,\lambda}}) - \mathcal{E}_z(\overrightarrow{h_\lambda}) &= \frac{1}{m}\sum_{i=1}^m \|\overrightarrow{y}_i - \overrightarrow{f_{z,\lambda}}(x_i)\|^2_{\mathbb{R}^{d+1}} - \frac{1}{m}\sum_{i=1}^m \|\overrightarrow{y}_i - \overrightarrow{h_\lambda}(x_i)\|^2_{\mathbb{R}^{d+1}} \\ &\ge \frac{1}{m}\sum_{i=1}^m \langle \overrightarrow{f_{z,\lambda}}(x_i) - \overrightarrow{h_\lambda}(x_i),\ -2(\overrightarrow{y}_i - \overrightarrow{h_\lambda}(x_i))\rangle_{\mathbb{R}^{d+1}}.\end{aligned}$$

Let $f_0^{(z,\lambda)}(x) = f^{(z,\lambda)}(x)$ and $h^{(\lambda)} = h_0^{(\lambda)}$. Then, by (1.5) we have

$$\begin{aligned}&\langle \overrightarrow{f_{z,\lambda}}(x_i) - \overrightarrow{h}_\lambda(x_i),\ -2(\overrightarrow{y}_i - \overrightarrow{h}_\lambda(x_i))\rangle_{\mathbb{R}^{d+1}} \\ &= -2\sum_{j=0}^d \Big(\frac{\partial f^{(z,\lambda)}(x_i)}{\partial x^j} - \frac{h^{(\lambda)}(x_i)}{\partial x^j}\Big) \times (y_i^{(j)} - h_j^{(\lambda)}(x_i)) \\ &= -2\sum_{j=0}^d \Big(f^{(z,\lambda)} - h^{(\lambda)},\ \frac{\partial K_{x_i}(\cdot)}{\partial x^j} \times (y_i^{(j)} - h_j^{(\lambda)}(x_i))\Big)_{\mathcal{H}_K} \\ &= \Big(f^{(z,\lambda)} - h^{(\lambda)}, -2\sum_{j=0}^d \frac{\partial K_{x_i}(\cdot)}{\partial x^j} \times (y_i^{(j)} - h_j^{(\lambda)}(x_i))\Big)_{\mathcal{H}_K} \\ &= (f^{(z,\lambda)} - h^{(\lambda)},\ -2\langle \nabla_x^* K_{x_i}(\cdot), (\overrightarrow{y}_i - \overrightarrow{h_\lambda}(x_i))\rangle_{\mathbb{R}^{d+1}})_{\mathcal{H}_K}.\end{aligned}$$

Therefore,

$$\mathcal{E}_z(\overrightarrow{f_{z,\lambda}}) - \mathcal{E}_z(\overrightarrow{h_\lambda}) \ge \Big(f^{(z,\lambda)} - h^{(\lambda)},\ -\frac{2}{m}\sum_{i=1}^m \langle \nabla_x^* K_{x_i}(\cdot),\ (\overrightarrow{y}_i - \overrightarrow{h_\lambda}(x_i))\rangle_{\mathbb{R}^{d+1}}\Big)_{\mathcal{H}_K}. \tag{2.17}$$

By the definition of $\overrightarrow{f_{z,\lambda}}$ we have

$$0 \ge (\mathcal{E}_z(\overrightarrow{f_{z,\lambda}}) + \lambda\|f^{(z,\lambda)}\|^2_{\mathcal{H}_K}) - (\mathcal{E}_z(\overrightarrow{h}_\lambda) + \lambda\|h^{(\lambda)}\|^2_{\mathcal{H}_K}). \tag{2.18}$$

Since $\mathcal{H}_K$ is a Hilbert space, the parallelogram law shows

$$\|f^{(z,\lambda)}\|^2_{\mathcal{H}_K} - \|h^{(\lambda)}\|^2_{\mathcal{H}_K} = (f^{(z,\lambda)} - h^{(\lambda)},\ 2h^{(\lambda)})_K + \|f^{(z,\lambda)} - h^{(\lambda)}\|^2_{\mathcal{H}_K}.$$

It follows by (2.17)–(2.18) that

$$\begin{aligned}0 \ge &\Big(f^{(z,\lambda)} - h^{(\lambda)},\ -\frac{2}{m}\sum_{i=1}^m \langle \nabla_x^* K_{x_i}(\cdot),\ (\overrightarrow{y}_i - \overrightarrow{h_\lambda}(x_i))\rangle_{\mathbb{R}^{d+1}}\Big)_{\mathcal{H}_K} \\ &+ (f^{(z,\lambda)} - h^{(\lambda)},\ 2\lambda h^{(\lambda)}) + \lambda\|f^{(z,\lambda)} - h^{(\lambda)}\|^2_{\mathcal{H}_K}.\end{aligned}$$

By (2.11) we have

$$\begin{aligned}0 \geq & \Big(f^{(z,\lambda)} - h^{(\lambda)},\ -\frac{2}{m}\sum_{i=1}^{m}\langle\nabla_x^* K_{x_i}(\cdot),\ (\overrightarrow{y}_i - \overrightarrow{h}_\lambda(x_i))\rangle_{\mathbb{R}^{d+1}} + 2\lambda h^{(\lambda)}\Big)_{\mathcal{H}_K} \\ & + \lambda\|f^{(z,\lambda)} - h^{(\lambda)}\|_{\mathcal{H}_K}^2 \\ = & \, 2\Big(f^{(z,\lambda)} - h^{(\lambda)},\ \int_Z \langle\nabla_x^* K_x(\cdot),\ \overrightarrow{y} - \overrightarrow{h^{(\lambda)}}(x)\rangle_{\mathbb{R}^{d+1}}\,\mathrm{d}\rho \\ & - \frac{1}{m}\sum_{i=1}^{m}\langle\nabla_x^* K_{x_i}(\cdot),\ (\overrightarrow{y}_i - \overrightarrow{h}_\lambda(x_i))\rangle_{\mathbb{R}^{d+1}}\Big)_{\mathcal{H}_K} \\ & + \lambda\|f^{(z,\lambda)} - h^{(\lambda)}\|_{\mathcal{H}_K}^2. \end{aligned} \tag{2.19}$$

(2.19) gives (2.16).

Lemma 2.7 *Let $\overrightarrow{h_\lambda}$ be defined as* (2.6) *and let $\overrightarrow{f_{z,\lambda}}$ be the solution of* (1.7)*. Then, there holds*

$$\|\overrightarrow{h_\lambda} - \overrightarrow{f_{z,\lambda}}\|_{L^2(\rho_X)}^2 \leq \frac{4}{\lambda}\, A(z)\ \ B(z), \tag{2.20}$$

where

$$A(z) = \Big\| \int_Z \langle\nabla_x^* K_x(\cdot),\ \overrightarrow{y} - \overrightarrow{h_\lambda}(x)\rangle_{\mathbb{R}^{d+1}}\,\mathrm{d}\rho - \frac{1}{m}\sum_{i=1}^{m}\langle\nabla_x^* K_{x_i}(\cdot),\ \overrightarrow{y}_i - \overrightarrow{h_\lambda}(x_i)\rangle_{\mathbb{R}^{d+1}}\Big\|_{\mathcal{H}_K}$$

and

$$B(z) = \Big\| \int_Z \langle\nabla_x^* K_x(\cdot),\ \overrightarrow{y} - \overrightarrow{f_{z,\lambda}}(x)\rangle_{\mathbb{R}^{d+1}}\,\mathrm{d}\rho - \frac{1}{m}\sum_{i=1}^{m}\langle\nabla_x^* K_{x_i}(\cdot),\ \overrightarrow{y}_i - \overrightarrow{f_{z,\lambda}}(x_i)\rangle_{\mathbb{R}^{d+1}}\Big\|_{\mathcal{H}_K}.$$

Proof By (2.9) we have

$$\begin{aligned} & \mathcal{E}_\rho(\overrightarrow{h_\lambda}) - \mathcal{E}_\rho(\overrightarrow{f_{z,\lambda}}) \\ = & \Big(\overrightarrow{h_\lambda} - \overrightarrow{f_{z,\lambda}},\ -2\int_Z \langle\nabla_x^* K_x(\cdot), \overrightarrow{y} - \overrightarrow{f_{z,\lambda}}(x)\rangle_{\mathbb{R}^{d+1}}\mathrm{d}\rho\Big)_{\mathcal{H}_K} + \|\overrightarrow{h_\lambda} - \overrightarrow{f_{z,\lambda}}\|_{L^2(\rho_X)}^2. \end{aligned} \tag{2.21}$$

The definitions of $\overrightarrow{h_\lambda}$ yields

$$0 \geq (\mathcal{E}_\rho(\overrightarrow{h}_\lambda) + \|h^\lambda\|_{\mathcal{H}_K}^2) - (\mathcal{E}_\rho(\overrightarrow{f}_{z,\lambda}) + \|f^{(z,\lambda)}\|_{\mathcal{H}_K}^2). \tag{2.22}$$

Further, by (2.21) and the equality

$$\|h^{(\lambda)}\|_{\mathcal{H}_K}^2 - \|f^{(z,\lambda)}\|_{\mathcal{H}_K}^2 = (h^{(\lambda)} - f^{(z,\lambda)},\ 2f^{(z,\lambda)})_K + \|h^{(\lambda)} - f^{(z,\lambda)}\|_{\mathcal{H}_K}^2,$$

we have

$$\begin{aligned} 0 \geq & \Big(h^{(\lambda)} - f^{(z,\lambda)},\ -2\int_Z \langle\nabla_x^* K_x(\cdot),\ \ \overrightarrow{y} - \overrightarrow{f_{z,\lambda}}(x)\rangle_{\mathbb{R}^{d+1}}\,\mathrm{d}\rho\Big)_{\mathcal{H}_K} \\ & + \lambda(h^{(\lambda)} - f^{(z,\lambda)},\ 2f^{(z,\lambda)})_{\mathcal{H}_K} + \lambda\,\|h^{(\lambda)} - f^{(z,\lambda)}\|_{\mathcal{H}_K}^2 + \Big\|\overrightarrow{h}_\lambda - \overrightarrow{f_{z,\lambda}}\Big\|_{L^2(\rho_X)}^2 \\ = & \, 2\Big(h^{(\lambda)} - f^{(z,\lambda)},\ -\int_Z \langle\nabla_x^* K_x(\cdot),\ \ \overrightarrow{y} - \overrightarrow{f_{z,\lambda}}(x)\rangle_{\mathbb{R}^{d+1}}\mathrm{d}\rho \\ & + \frac{1}{m}\sum_{i=1}^{m}\langle\nabla_x^* K_{x_i}(\cdot),\ \ \overrightarrow{y}_i - \overrightarrow{f_{z,\lambda}}(x_i)\rangle_{\mathbb{R}^{d+1}}\Big)_{\mathcal{H}_K} + \lambda\|h^{(\lambda)} - f^{(z,\lambda)}\|_{\mathcal{H}_K}^2 \\ & + \Big\|\overrightarrow{h}_\lambda - \overrightarrow{f_{z,\lambda}}\Big\|_{L^2(\rho_X)}^2, \end{aligned} \tag{2.23}$$

where we have used (2.12). By Cauchy's inequality we have

$$\begin{aligned}&\lambda\|h^{(\lambda)}-f^{(z,\lambda)}\|_{\mathcal{H}_K}^2+\|\overrightarrow{h_\lambda}-\overrightarrow{f_{z,\lambda}}\|_{L^2(\rho_X)}^2\\&\le 2\,\|h^{(\lambda)}-f^{(z,\lambda)}\|_{\mathcal{H}_K}\times\Big\|\int_Z\langle\nabla_x^*K_x(\cdot),\ \overrightarrow{y}-\overrightarrow{f_{z,\lambda}}(x)\rangle_{\mathbb{R}^{d+1}}\mathrm{d}\rho\\&-\frac{1}{m}\sum_{i=1}^m\langle\nabla_x^*K_{x_i}(\cdot),\ \overrightarrow{y}_i-\overrightarrow{f_{z,\lambda}}(x_i)\rangle_{\mathbb{R}^{d+1}}\Big\|_{\mathcal{H}_K}.\end{aligned}$$

Above inequality leads to

$$\lambda\|h^{(\lambda)}-f^{(z,\lambda)}\|_{\mathcal{H}_K}+\frac{\|\overrightarrow{h_\lambda}-\overrightarrow{f_{z,\lambda}}\|_{L^2(\rho_X)}^2}{\|h^{(\lambda)}-f^{(z,\lambda)}\|_{\mathcal{H}_K}}\le 2\ B(z).$$

It follows that

$$\frac{\|\overrightarrow{h_\lambda}-\overrightarrow{f_{z,\lambda}}\|_{L^2(\rho_X)}^2}{\|h^{(\lambda)}-f^{(z,\lambda)}\|_{\mathcal{H}_K}}\le 2\ B(z).\tag{2.24}$$

By (2.16) and (2.24) we have (2.20).

Lemma 2.8 (see [19]) *Let $(H,\|\cdot\|)$ be a Hilbert space and ξ be a random variable on (Z,ρ) with values in H. Assume that $\|\xi\|_H\le\widetilde{M}<+\infty$ almost surely. Let $\{z_i\}_{i=1}^m$ be independent samples drawers of ρ. For any $0<\delta<1$, with confidence $1-\delta$,*

$$\Big\|\frac{1}{m}\sum_{i=1}^m\xi(z_i)-E(\xi)\Big\|_H\le\frac{2\widetilde{M}\ \log(\frac{2}{\delta})}{\sqrt{m}}.\tag{2.25}$$

Lemma 2.9 (see [6]) *Let $\mathcal{F}$ be a family of functions from a probability space Z to $\Re$ and $d(\cdot,\cdot)$ be a distance on $\mathcal{F}$. Let $\mathcal{U}\subset Z$ be of full measure and constants $H,p>0$ such that*

(i) *$|\xi(z)|\le H$ for all $\xi\in\mathcal{F}$ and all $z\in\mathcal{U}$, and*

(ii) *$|L_z(\xi_1)-L_z(\xi_2)|\le p\ d(\xi_1,\ \xi_2)$ for all $\xi_1,\xi_2\in\mathcal{F}$ and all $z\in\mathcal{U}^m$, where*

$$L_z(\xi)=\int_Z\xi(z)-\frac{1}{m}\sum_{i=1}^m\xi(z_i).$$

Then, for all $\epsilon>0$,

$$\mathrm{Prob}_{z\in Z^m}\Big\{\sup_{\xi\in\mathcal{F}}|L_z(\xi)|\le\epsilon\Big\}\ge 1-\mathcal{N}\Big(\mathcal{F},\ \frac{\epsilon}{2p}\Big)\times 2\exp\Big(-\frac{m\ \epsilon^2}{8\ H^2}\Big).\tag{2.26}$$

Lemma 2.10 *Under the conditions of Theorem* 1.1, *we have the following two estimates:*

(1) *For any $\delta\in(0,1)$, with confidence $1-\frac{\delta}{2}$, we have*

$$A(z)\le\frac{16\ k_1^2\ d\ \sqrt{D(\overrightarrow{f_\rho},\ \lambda)}}{\sqrt{m\lambda}}\times\log\frac{4}{\delta}.\tag{2.27}$$

(2) *For any $\delta\in(0,1)$ and $0<\delta<\frac{4}{e^{c_s}}$, with confidence $1-\frac{\delta}{2}$, we have*

$$B(z)\le\frac{4k_1^2\ (d+1)^2M}{\sqrt{\lambda}}\times\Big(\frac{1}{\sqrt{m}}+\frac{1}{\sqrt[2+s]{m}}\Big)\log\frac{4}{\delta}.\tag{2.28}$$

Proof of (2.27) Take

$$\xi(x, \overrightarrow{y}, \cdot) = \langle \nabla_x^* K_x(\cdot),\ \overrightarrow{y} - \overrightarrow{h_\lambda}(x) \rangle_{\mathbb{R}^{d+1}}, \quad (x, \overrightarrow{y}) \in X \times Y.$$

Since

$$\nabla_x^* K_x(\cdot) = \Big(\frac{\partial K_x(\cdot)}{\partial x^0},\ \frac{\partial K_x(\cdot)}{\partial x^1}, \cdots, \frac{\partial K_x(\cdot)}{\partial x^d} \Big)^{\mathrm{T}}, \quad \frac{\partial K_x(\cdot)}{\partial x^0} = K_x(\cdot),$$

we have

$$\langle \nabla_x^* K_x(\cdot),\ \overrightarrow{y} - \overrightarrow{h_\lambda}(x) \rangle_{\mathbb{R}^{d+1}} = \sum_{i=0}^{d} \frac{\partial K_x(\cdot)}{\partial x^i} \times (y^i - h_i^{(\lambda)}(x)). \tag{2.29}$$

By (1.5), we have

$$\begin{aligned}
\|\xi(x, \overrightarrow{y}, \cdot)\|_{\mathcal{H}_K}^2 &= (\xi(x, \overrightarrow{y}, \cdot),\ \xi(x, \overrightarrow{y}, \cdot))_{\mathcal{H}_K} \\
&= \sum_{i,j=0}^{d} \Big(\frac{\partial K_x(\cdot)}{\partial x^i},\ \frac{\partial K_x(\cdot)}{\partial x^j} \Big)_{\mathcal{H}_{K,\rho_X}} \times (y^i - h_i^{(\lambda)}(x)) \times (y^j - h_j^{(\lambda)}(x)) \\
&= \sum_{i,j=0}^{d} \frac{\partial^2 K_x(x)}{\partial x^i \partial x^j} \times (y^i - h_i^{(\lambda)}(x)) \times (y^j - h_j^{(\lambda)}(x)) \\
&\le k_1^2\ d^2\ \sum_{i=0}^{d} |y^i - h_i^{(\lambda)}(x)|^2 \\
&= k_1^2\ d^2\ \|\overrightarrow{y} - \overrightarrow{h_\lambda}(x)\|_{\mathbb{R}^{d+1}}^2 \\
&\le 4\ k_1^2\ d^2\ (M^2 + \|\overrightarrow{h_\lambda}(x)\|_{\mathbb{R}^{d+1}}^2).
\end{aligned} \tag{2.30}$$

Furthermore, by (2.14) we have

$$\begin{aligned}
\|\overrightarrow{h_\lambda}(x)\|_{\mathbb{R}^{d+1}}^2 &= \sum_{i=0}^{d} \Big| \frac{\partial h^{(\lambda)}(x)}{\partial x^i} \Big|^2 \\
&\le k_1^2\ d\ \|h^{(\lambda)}\|_{\mathcal{H}_K}^2 \le \frac{k_1^2\ d\ D(\overrightarrow{f_\rho},\ \lambda)}{\lambda}.
\end{aligned} \tag{2.31}$$

Therefore, if $\lambda \le \frac{k_1^2 d\ D(\overrightarrow{f_\rho}, \lambda)}{M^2}$, then

$$\begin{aligned}
\|\xi(x, \overrightarrow{y}, \cdot)\|_{\mathcal{H}_K}^2 &\le 4k_1^2 \Big(M^2 + \frac{k_1^2 d\ D(\overrightarrow{f_\rho},\ \lambda)}{\lambda} \Big) \\
&\le \frac{8k_1^4 d\ D(\overrightarrow{f_\rho},\ \lambda)}{\lambda}.
\end{aligned} \tag{2.32}$$

By (2.32) and (2.25), we have (2.27).

Proof of (2.28) By the definition of $\|\cdot\|_{\mathcal{H}_K}$, we have

$$\begin{aligned} B = & \sup_{\|g\|_{\mathcal{H}_K}\le 1} \Big| \Big(g, \int_Z \langle \nabla_x^* K_x(\cdot), \overrightarrow{y} - \overrightarrow{f_{z,\lambda}}(x)\rangle_{\mathbb{R}^{d+1}}\, \mathrm{d}\rho \\ & - \frac{1}{m}\sum_{i=1}^m \langle \nabla_x^* K_{x_i}(\cdot), \overrightarrow{y}_i - \overrightarrow{f_{z,\lambda}}(x_i)\rangle_{\mathbb{R}^{d+1}}\Big)_{\mathcal{H}_K}\Big| \\ = & \sup_{\|g\|_{\mathcal{H}_K}\le 1} \Big| \int_Z \Big(g, \langle \nabla_x^* K_x(\cdot), \overrightarrow{y} - \overrightarrow{f_{z,\lambda}}(x)\rangle_{\mathbb{R}^{d+1}}\, \mathrm{d}\rho\Big)_{\mathcal{H}_K} \\ & - \Big(g, \frac{1}{m}\sum_{i=1}^m \langle \nabla_x^* K_{x_i}(\cdot), \overrightarrow{y}_i - \overrightarrow{f_{z,\lambda}}(x_i)\rangle_{\mathbb{R}^{d+1}}\Big)_{\mathcal{H}_K}\Big|. \end{aligned}$$

By (2.10), we have

$$B = \sup_{\|g\|_{\mathcal{H}_K}\le 1} \Big| \int_Z \langle \overrightarrow{g}(x), \overrightarrow{y} - \overrightarrow{f}_{z,\lambda}(x)\rangle_{\mathbb{R}^{d+1}} \mathrm{d}\rho - \frac{1}{m}\sum_{i=1}^m \langle \overrightarrow{g}(x_i), \overrightarrow{y}_i - \overrightarrow{f}_{z,\lambda}(x_i)\rangle_{\mathbb{R}^{d+1}}\Big|.$$

Define

$$\eta(x, \overrightarrow{y}) = \langle \overrightarrow{g}(x), \overrightarrow{y} - \overrightarrow{f}_{z,\lambda}(x)\rangle_{\mathbb{R}^{d+1}}.$$

Then

$$|\eta(x, \overrightarrow{y})| \le \|\overrightarrow{g}(x)\|_{\mathbb{R}^{d+1}} \times \|\overrightarrow{y} - \overrightarrow{f}_{z,\lambda}(x)\|_{\mathbb{R}^{d+1}}. \tag{2.33}$$

By (1.6), we have

$$\begin{aligned} \|\overrightarrow{g}(x)\|_{\mathbb{R}^{d+1}} &= \Big(\sum_{i=0}^d |g_i(x)|^2\Big)^{\frac{1}{2}} \\ &= \Big(\sum_{i=0}^d \Big|\frac{\partial g(x)}{\partial x^i}\Big|^2\Big)^{\frac{1}{2}} \\ &\le (d+1)k_1\, \|g\|_{\mathcal{H}_K} \\ &\le (d+1)\, k_1, \end{aligned} \tag{2.34}$$

and by (2.15), we have

$$\begin{aligned} \|\overrightarrow{y} - \overrightarrow{f}_{z,\lambda}(x)\|_{\mathbb{R}^{d+1}} &\le \|\overrightarrow{y}\|_{\mathbb{R}^{d+1}} + \|\overrightarrow{f}_{z,\lambda}(x)\|_{\mathbb{R}^{d+1}} \\ &\le (d+1)(M + k_1\|f^{(z,\lambda)}\|_{\mathcal{H}_K}) \\ &\le (d+1)M\Big(1 + \frac{k_1}{\sqrt{\lambda}}\Big). \end{aligned}$$

Therefore,

$$\begin{aligned} |\eta(x, \overrightarrow{y})| &\le (d+1)^2 k_1 M\Big(1 + \frac{k_1}{\sqrt{\lambda}}\Big) \\ &\le H = \frac{2(d+1)^2 k_1^2 M}{\sqrt{\lambda}} \end{aligned} \tag{2.35}$$

if $\lambda \le k_1^2$.

Take

$$\mathcal{F} = \{\eta(x, \overrightarrow{y}) \in C(Z) : \ |\eta(x, \overrightarrow{y})| \le H\}.$$

Then,

$$\begin{aligned} B &= \sup_{\eta\in\mathcal{F}} \Big| \int_Z \eta(x, \overrightarrow{y})\mathrm{d}\rho - \sum_{i=1}^m \eta(x_i, \overrightarrow{y}_i)\Big| \\ &= \sup_{\eta\in\mathcal{F}} |\mathcal{L}_z(\eta)|, \end{aligned} \tag{2.36}$$

where

$$\mathcal{L}_z(\eta) = \int_Z \eta(x, \overrightarrow{y})\mathrm{d}\rho - \frac{1}{m}\sum_{i=1}^m \eta(x_i, \overrightarrow{y}_i), \qquad z = (x, \overrightarrow{y})$$

satisfies the inequality

$$|\mathcal{L}_z(\eta_1) - \mathcal{L}_z(\eta_2)| \le 2\|\eta_1 - \eta_2\|_{C(Z)}, \qquad \eta_1 \in \mathcal{F},\ \eta_2 \in \mathcal{F}.$$

By Lemma 2.9, we have for any $\varepsilon > 0$

$$\begin{aligned} \mathrm{Prob}_{z\in Z^m}(|\mathcal{L}_z(\eta)| \le \varepsilon) &\ge 1 - \mathcal{N}\Big(\mathcal{F}, \frac{\varepsilon}{4}\Big)2\exp\Big(-\frac{m\varepsilon^2}{8H^2}\Big) \\ &\ge 1 - 2\exp\Big(c_s\Big(\frac{4H}{\varepsilon}\Big)^s - \frac{m\varepsilon^2}{8H^2}\Big). \end{aligned}$$

Take

$$2\exp\Big(c_s\Big(\frac{4H}{\varepsilon}\Big)^s - \frac{m\varepsilon^2}{8H^2}\Big) = \delta.$$

Then, we have

$$\varepsilon^{2+s} - \frac{8H^2\varepsilon^2}{m}\log\frac{4}{\delta} - \frac{8\times 4^s c_s H^{2+s}}{m} = 0. \tag{2.37}$$

Notice that there is the famous lemma (see [26]).

Let $c_1 > 0,\ c_2 > 0$ and $u > t > 0$. Then the equation

$$x^u - c_1 x^t - c_2 = 0$$

has a unique positive zero x^*. In addition,

$$x^* \le \max\{(2\,c_1)^{\frac{1}{u-t}},\ (2\,c_2)^{\frac{1}{u}}\}. \tag{2.38}$$

By (2.37)–(2.38), we have

$$\begin{aligned} \varepsilon &\le \sqrt{\frac{16H^2}{m}\log\frac{4}{\delta}} + \sqrt[2+s]{\frac{16\times 4^s c_s H^{2+s}}{m}} \\ &= \frac{8(d+1)^2 k_1^2}{\sqrt{\lambda}}\Big(\sqrt{\frac{1}{m}\log\frac{4}{\delta}} + \sqrt[2+s]{\frac{c_s}{m}}\Big). \end{aligned} \tag{2.39}$$

By (2.36) and (2.39), we have (2.28).

Lemma 2.11 *For any $\delta \in (0,1)$, with confidence $1-\delta$, we have*

$$\begin{aligned} \|\overrightarrow{f_{z,\lambda}} - \overrightarrow{h_\lambda}\|^2_{L^2(\rho_X)} \le{}& \frac{512\ k_1^4\ (d+1)^3\ M\ \sqrt{D(\overrightarrow{f_\rho},\ \lambda)}}{\lambda^2\sqrt{m}} \\ &\times \Big(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt[2+s]{m}}\Big) \times \Big(\log\frac{4}{\delta}\Big)^2. \end{aligned} \tag{2.40}$$

Proof By (2.20), we have

$$\frac{\lambda}{4}\|\overrightarrow{h_\lambda} - \overrightarrow{f_{z,\lambda}}\|^2_{L^2(\rho_X)} \le A(z)\ B(z).$$

Then,

$$\begin{aligned}
&\Big\{z \in Z^m :\ \frac{\lambda}{4}\|\overrightarrow{f_{z,\lambda}} - \overrightarrow{h_\lambda}\|^2_{L^2(\rho_X)} \le \frac{128\ k_1^4\ (d+1)^3\ M\ \sqrt{D(\overrightarrow{f_\rho},\lambda)}}{m\lambda} \times \Big(\log\frac{4}{\delta}\Big)^2\Big\} \\
\supset&\Big\{z \in Z^m :\ A(z) \le \frac{16\ k_1^2\ d\sqrt{D(\overrightarrow{f_\rho},\lambda)}}{\sqrt{m\lambda}} \times \log\frac{4}{\delta}\Big\} \\
\cap&\Big\{z \in Z^m :\ B(z) \le \frac{8\ k_1^2\ (d+1)^2\ M}{\sqrt{\lambda}} \times \Big(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt[s]{m}^{2+s}}\Big) \times \log\frac{4}{\delta}\Big\}. \qquad (2.41)
\end{aligned}$$

By (2.27)–(2.28) and (2.41), we have (2.40).

Proof of (1.9) By the definition of $\overrightarrow{h_\lambda}$, (2.3)–(2.4) we have

$$\begin{aligned}
\|\overrightarrow{f_\rho} - \overrightarrow{f_{z,\lambda}}\|^2_{L^2(\rho_X)} &\le 4\|\overrightarrow{h_\lambda} - \overrightarrow{f_{z,\lambda}}\|^2_{L^2(\rho_X)} + 4\|\overrightarrow{f_\rho} - \overrightarrow{h_\lambda}\|^2_{L^2(\rho_X)} \\
&\le 4\|\overrightarrow{h_\lambda} - \overrightarrow{f_{z,\lambda}}\|^2_{L^2(\rho_X)} + 4(\|\overrightarrow{f_\rho} - \overrightarrow{h_\lambda}\|^2_{L^2(\rho_X)} + \lambda\|h_\lambda\|^2_{\mathcal{H}_K}) \\
&= 4\|\overrightarrow{h_\lambda} - \overrightarrow{f_{z,\lambda}}\|^2_{L^2(\rho_X)} + 4(\mathcal{E}_\rho(\overrightarrow{h_\lambda}) - \mathcal{E}(\overrightarrow{f}_\rho) + \lambda\ \|h_\lambda\|^2_{\mathcal{H}_K}) \\
&= 4\|\overrightarrow{h}_\lambda - \overrightarrow{f_{z,\lambda}}\|^2_{L^2(\rho_X)} + 4D(\overrightarrow{f_\rho},\ \lambda). \qquad (2.42)
\end{aligned}$$

(2.40) and (2.42) give (1.9).

References

[1] Zhou, D. X., Derivative reproducing properties for kernel methods in learning theory, *Journal of Computational and Applied Mathematics*, **220**, 2008, 456–463.

[2] Shi, L., Guo, X. and Zhou, D. X., Hérmite learning with gradient data, *Journal of Computational and Applied Mathematics*, **233**, 2010, 3046–3056.

[3] Mukherjee, S., Wu, Q. and Zhou, D. X., Learning gradients on manifolds, *Bernoulli*, **16**(1), 2010, 181–207.

[4] Ying, Y. M., Wu, Q. and Campbell, C., Learning the coordinate gradients, *Advances in Computational Mathematics*, **37**(3), 2012, 355–378.

[5] Cucker, F. and Smale, F., On the mathematical foundations of learning theory, *Bull. Amer. Math.*, **39**, 2001, 1–49.

[6] Cucker, F. and Zhou, D. X., Learning Theory: An Approximation Theory Viewpoint, Cambridge University Press, New York, 2007.

[7] Aronszajn, N., Theory of reproducing kernels, *Transactions of the American Mathematical Society*, **68**, 1950, 337–404.

[8] Belkin, M., Niyogi, P. and Sindhwani, V., Manifold regularization: A geometric framework for learning from labeled and unlabeled examples, *Journal of Machine Learning Research*, **7**, 2006, 2399–2434.

[9] Smale, S. and Zhou, D. X., Estimating the approximation error in learning theory, *Analysis and Applications*, **1**(1), 2003, 17–41.

[10] Solnon, M., Arlot, S. and Bach, F., Multi-task regression using minimal penalities, *Journal of Machine Learning Research*, **13**, 2012, 2773–2812.

[11] Caponnetto, A., Micchelli, C. A., Pontil, M. and Ying, Y. M., Uinversal multi-task kernels, *Journal of Machine Learning Research*, **9**, 2008, 1615–1646.

[12] Wang, J. Y., Bensmail, H. and Gao, X., Feature selection and multi-kernel learning for sparse representation on a manifold, *Neural Networks*, **51**, 2014, 9–16.

[13] Evgeniou, T., Micchelli, C. A. and Pontil, M., Learning multiple tasks with kernel methods, *Journal of Machine Learning Research*, **6**, 2005, 615–637.

[14] Zhang, H. Z. and Zhang, J., Vector-valued reproducing kernel Banach spaces with applications to multi-task learning, *Journal of Complexity*, **29**(2), 2013, 195–215.

[15] Jordão, J. and Menegatto, V. A., Reproducing properties of differentiable Mercer-like kernels on the sphere, *Numerical Functional Analysis and Optimization*, **33**(10), 2012, 1221–1243.

[16] Ferreira, J. C. and Menegatto, V. A., Reproducing properties of differentiable Mercer-like kernels, *Mathematische Nachrichten*, **285**(8–9), 2012, 959–973.

[17] Jordão, T. and Menegatto, V. A., Weighted Fourier-Laplace transforms in reproducing kernel Hilbert spaces on the sphere, *Journal of Mathematical Analysis and Applications*, **411**, 2014, 732–741.

[18] Sun, H. W. and Wu, Q., A note on application of integral operator in learning, *Applied and Computational Harmonic Analysis*, **26**, 2009, 416–421.

[19] Smale, S. and Zhou, D. X., Learning theory estimates via integral operators and their applications, *Constructive Approximation*, **26**, 2007, 153–172.

[20] Sheng, B. H. and Ye, P. X., The learning rates of regularized regression based on reproducing kernel Banach spaces, Abstract and Applied Analysis, 2013, Article ID 694181, 10 pages, http://dx. doi .org/10.1155/2013/694181.

[21] Sheng, B. H., The convergence rates of Shannon sampling learning algorithms, *Sciences in China: Mathematics*, **55**(6), 2012, 1243–1256.

[22] Christmann, A. and Steinwart, I., Consistency and robustness of kernel-based regression in convex risk minimization, *Bernoulli*, **13**(3), 2007, 799–819.

[23] Steinwart, I., Sparseness of support vector machines, *Journal of Machine Learning Research*, **4**, 2003, 1071–1105.

[24] Sheng, B. H., Xiang, D. H. and Ye, P. X., Convergence rate of semi-supervised gradient learning algorithms, *International Journal of Wavelets, Multiresolution and Information Processing*, **13**(4), 2015, 1550021 (26 pages).

[25] Bauschke, H. H. and Combettes, P. L., Convex Analysis and Monotone Operator Theory in Hilbert Spaces, Springer-Verlag, New York, 2010.

[26] Cucker, F. and Smale, S., Best choices for regularization parameters in learning theory: On the bias-variance problem, *Foundation of Computational Mathematics*, **2**, 2002, 413–428.