The learning rate of vector-valued ranking with least square loss

LIU HUANXIANG
School of Statistics and Mathematics,
Zhejiang Gongshang University,
Hangzhou, 310018,
P.R.China
e-mail: lhxiang@usx.edu.cn

SHENG BAOHUAI*
Department of Mathematics,
Shaoxing University
Shaoxing, Zhejiang, 312000,
P.R. China
e-mail: bhsheng@usx.edu.cn

YE PEIXIN
School of Mathematics and LPMC
Nankai University
Tianjin 300071
P. R. China
e-mail: yepx@nankai.edu.cn

Abstract: In the present paper, we give an investigation on the quantitative convergence analysis of the kernel regularized vector ranking with least square loss. We present with Gâteaux derivative the qualitative relation between the solution and the hiding distribution and quantitatively show the robustness for the solution. Finally, we provide a learning rate in terms of the approximation ability and capacity of the involved vector-valued RKHS.

Key Words: Vector ranking, least square loss, Gâteaux derivative, vector-valued RKHS

1 Introduction

Ranking is a new learning problem focusing on relative ranking of objects on the basis of their observed features (see e.g.[2, 25]). It has been widely used in many fields as information retrieval, banking, quality control or survival analysis,et al (see [1, 13, 14, 17, 26, 32, 35, 40]). Recently, ranking theory is unified with the machine learning and the regularized kernel ranking is formed. For example, the extreme ranking learning and l_2-coefficient regularized ranking are defined in [15] and [12] respectively. The semi-supervised ranking is considered in [24]. The performance of the kernel regularized bipartite ranking with convex losses is discussed in [18].

A general ranking setting defined by [2] is:

The learner is given examples of instances labeled by real numbers, and the goal is to learn a ranking in which instances labeled by larger numbers are ranked higher than instances labeled by smaller numbers.

Let $z = \{(x_i,\ y_i)\}_{i=1}^m \in Z^m = (X \times R)^m$ be some observations with X being a given compact set and $Y = R$ the real number set. Then the aim of ranking is to find from the observations z a function $y = f_z(x) :\ X \to R$,called a ranking rule, that x is to be ranked preferred over x' if $y - y' \geq 0$ and lower than x' if $y - y' < 0$. Ranking with least square loss is to regress the differences $y_i - y_j$ with $f(x_i) - f(x_j)$ (see [21, 22, 23, 26, 34]). In this case, the empirical loss function $l_{sq} : \mathcal{H} \times Z \times Z \to R^+ \cup \{0\}$ is defined as

$$\begin{aligned} & l_{sq}(f,\ (x_i, y_i)) \\ = & \Big(|y_i - y_j| - sgn(y_i - y_j)(f(x_i) - f(x_j))\Big)^2 \\ = & \Big((y_i - y_j) - (f(x_i) - f(x_j))\Big)^2, \qquad (1) \end{aligned}$$

where $\mathcal{H}$ is a function space with metric $\|\cdot\|_{\mathcal{H}}$ and is called hypothesis space.

The corresponding ranking scheme takes the form (see [2, 11])

$$f_z^* := \arg\min_{f \in \mathcal{H}} \Big(\mathcal{E}_z^*(f) + \lambda\ \|f\|_{\mathcal{H}}^2\Big), \qquad (2)$$

where

$$\begin{aligned} \mathcal{E}_z^*(f) = & \frac{1}{m(m-1)} \sum_{i,j=1, i \neq j}^{m} \\ & \times [(y_i - y_j) - (f(x_i) - f(x_j))]^2 \end{aligned}$$

[*] Corresponding author

is the empirical error and $\lambda > 0$ are the regularization parameters. The expected error $\mathcal{E}(f)$ is

$$\begin{aligned}\mathcal{E}(f) &= \int_Z \int_Z \Big[(y-y') - (f(x) - f(x'))\Big]^2 \\ &\quad \times d\rho(x,y)\, d\rho(x',y').\end{aligned}$$

To show the predictive ability of (2), one need to compare the risk of f_z with the risk of the best rule. In learning theory, we use the probabilistic inequality of the following form (see [26]):

$$P\Big(\mathcal{E}(f) - \mathcal{E}(f_\rho^*) \leq \eta\Big) \geq 1 - \alpha, \tag{3}$$

to show the error $\mathcal{E}(f) - \mathcal{E}(f_\rho^*)$ with given confidence α, where $\eta > 0$ is some small numbers which depends on the level α,the number m of elements in the sample and the hypothesis space $\mathcal{H}$, but it is independent of the unknown distribution ρ. In case of the least square loss we choose

$$f_\rho^*(x) = \int_Y y\, d\rho(y|x)$$

as the best rule since

$$f_\rho^* = \arg\min_{f \in \mathcal{F}} \mathcal{E}(f),$$

where $\mathcal{F}$ is the set of all the measurable functions on X.

On the other hand, we find that the vector ranking with respect to the relevance vector machine has been proposed (see e.g.[10, 33, 36]). To extend above mentioned ranking theory to general vector form, we first consider the Euclidean space R^d. Assume $S = \{(x_1, y_1), \cdots, (x_m, y_m)\}$ is a finite sequence of labeled training examples,where x_i are instances in X and y_i are vector-valued labels in $Y = [0,\ M]^d \subset R^d$.

For the labels $y_j = (y_j^1,\ \cdots,\ y_j^d)^\top \in R^d$, $y_i = (y_i^1,\ \cdots,\ y_i^d)^\top \in R^d$, we say $y_i \geq y_j$ if and only if for all $k = 1, 2, \cdots, d$ we have $y_i^k \geq y_j^k$. We say x_i is ranked over x_j if $y_i \geq y_j$ and lower than x_j if $y_i < y_j$. The case of neither $y_i \geq y_j$ nor $y_i < y_j$ indicates no ranking preference between the two input instances.

Assume $f_z = (f_z^1, \cdots, f_z^d)$ is a predictive rule determined with the instances z. Then, according to above ranking theory in real valued function, we need to consider d least square single variable rankings:

$$\begin{aligned}& l_{sq}(f^k, (x_i, y_i)) \\ =\ & \Big((y_i^k - y_j^k) - (f^k(x_i) - f^k(x_j))\Big)^2, \\ & k = 1, 2, \cdots, d.\end{aligned}$$

The total number is $m(m-1)$, therefore, there are following weight means

$$\begin{aligned}& \frac{1}{m(m-1)} \sum_{i,j=1, i\neq j}^m \sum_{k=1}^d l_{sq}(f^k, (x_i, y_i)) \\ =\ & \frac{1}{m(m-1)} \sum_{i,j=1, i\neq j}^m \sum_{k=1}^d \\ & \times \Big((y_i^k - y_j^k) - (f^k(x_i) - f^k(x_j))\Big)^2 \\ =\ & \frac{1}{m(m-1)} \sum_{i,j=1, i\neq j}^m \\ & \times \|(y_i - y_j) - (f(x_i) - f(x_j))\|_{R^d}^2, \end{aligned} \tag{4}$$

where $\|\cdot\|_{R^d}$ is the Euclidean norm. (4) is the vector R^d-valued ranking corresponding to the strong order $\geq$, extending the R^d-norm $\|\cdot\|_{R^d}$ to the general vector norm, we shall have a kind of vector-valued ranking.

1.1 Vector-valued ranking

Let Λ be a Hilbert space and X be a set. We call $\mathcal{H}$ an Λ-valued Hilbert space on X if

$\mathcal{H}$ is a Hilbert space of functions from X to Λ and $\|f\|_{\mathcal{H}} = 0$ if and only if $f(x) = 0$ for all $x \in X$.

For two Banach spaces Λ_1 and Λ_2 we denote by $\mathcal{M}(\Lambda_1,\ \Lambda_2)$ the set of all the bounded operators from Λ_1 to Λ_2 and $\mathcal{L}(\Lambda_1, \Lambda_2)$ the subset of $\mathcal{M}(\Lambda_1, \Lambda_2)$ of those bounded operators that are also linear.

If $\Lambda_1 = \Lambda_2 = \Lambda$, then, $\mathcal{M}(\Lambda_1, \Lambda_2)$ is abbreviated as $\mathcal{M}(\Lambda)$. For each $T \in \mathcal{M}(\Lambda_1, \Lambda_2)$,we denote by $\|T\|_{\mathcal{M}(\Lambda_1,\ \Lambda_2)}$ the greatest lower bound of all the nonnegative constants α such that

$$\|Tu\|_{\Lambda_2} \leq \alpha \|u\|_{\Lambda_1}, \qquad \text{for all } u \in \Lambda_1.$$

When T is also linear, this quantity equals the operator norm $\|T\|_{\mathcal{L}(\Lambda_1,\ \Lambda_2)}$.

We call an Λ-valued Hilbert space $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ an RKHS on X if for all $x \in X$ there exists a positive constant C_x such that

$$\|f(x)\|_\Lambda \leq C_x \|f\|_{\mathcal{H}}, \qquad \forall f \in \mathcal{H}.$$

A bivariate symmetric function $K(x, y) : X \times X \to \mathcal{M}(\Lambda)$ is called an Λ-valued positive definite kernel if for all positive integers $N, x_1, \cdots, x_N \in X$ and $c_1, \cdots, c_N \in \mathcal{C}$,

$$\sum_{i,j=1}^N c_i\, c_j \langle K(x_j, x_i)v,\ v\rangle_\Lambda \geq 0, \qquad \forall v \in \Lambda.$$

Let $\mathcal{H}$ be an Λ-valued RKHS. Then, there exists a function $K(x, y)$ from $X \times X$ to $\mathcal{M}(\Lambda)$ such that for all $f \in \mathcal{H},\ x \in X$ and $\xi \in \Lambda$ there holds

$$\langle f(x),\ \xi\rangle_\Lambda = (f,\ \ K(\cdot,\ x)\xi)_{\mathcal{H}}. \tag{5}$$

If $\Lambda = R$, then, the R-valued RKHS is the usual RKHS (see [4]).

By [8, 9] we know that for a given Λ-valued positive definite kernel $K(x,\ y)$, there is uniquely an Λ-valued RKHS $\mathcal{H}$ on X with reproducing kernel $K(x,\ y)$.

Throughout the paper, we assume $C(X, \Lambda)$ is the set of all the functions $f(x)\ :\quad X \to \Lambda$ such that $\|f(x)\|_\Lambda$ is continuous on X, $K(x, y) :\ X \times X \to \Lambda$ is continuous on $X \times Y$ and $K(x, x)$ is a compact operator for all $x \in X$. Then, the inclusion $I_K : \mathcal{H} \to C(X, \Lambda)$ is compact (see the Proposition 13 of [9]) and there holds the inequality

$$\|f(x)\|_\Lambda \le k\|f\|_{\mathcal{H}}, \qquad x \in X,\ f \in \mathcal{H}, \tag{6}$$

where $k = \max\limits_{x\in X} \sqrt{\Big\|K(x,\ x)\Big\|_{\mathcal{L}(\Lambda)}}$.

Let $M_\Lambda \subset \Lambda$ be a bounded set with upper bound M, i.e., $M_\Lambda = \{y \in \Lambda : \|y\|_\Lambda \le M\}$. Let $\rho(x, y) = \rho(y|x)\rho_X(x)$ be an Λ-valued distribution on $Z = X \times M_\Lambda$, according to which the samples $z = \{(x_i, y_i)\}_{i=1}^m$ are drawn independently. Then, the vector ranking based on an Λ-valued RKHS $\mathcal{H}$ and the least square loss is

$$f_{z,\lambda} = arg \min_{f\in\mathcal{H}} \mathcal{E}_z(f) + \lambda\|f\|^2_{\mathcal{H}}, \tag{7}$$

where $\mathcal{E}_z(f)$ is the empirical error defined as

$$\begin{aligned}\mathcal{E}_z(f) \quad &= \quad \frac{1}{m(m-1)} \sum_{i,j=1,i\ne j}^{m} \\ &\quad \times\|(y_i - y_j) - (f(x_i) - f(x_j))\|^2_\Lambda. \quad (8)\end{aligned}$$

When $\Lambda = R^d$, we have the R^d-valued empirical error (4). Therefore, framework (7) is the vector generalization of (2). The purpose of the present paper is to give a quantitatively performance analysis for (7).

1.2 The learning rate

Define the expected error corresponding to (7) as

$$\begin{aligned}\mathcal{E}_\rho(f) \quad &= \quad \int_Z\int_Z \Big\|(y - f(x)) - (y' - f(x'))\Big\|^2_\Lambda \\ &\quad \times\ d\rho(x, y)\ d\rho(x',\ y').\end{aligned}$$

and

$$\mathcal{G} = \{f \in \mathcal{F} :\ \ f = \arg \inf_{f\in\mathcal{F}} \mathcal{E}_\rho(f)\},$$

where $\mathcal{F}$ is the class of all Λ-valued ρ_X measurable functions (see from[31]). Then, by Lemma 2.1 in Section 2 we know the Λ-valued regression function $f_\rho(x)$ (defined as in [7])

$$f_\rho(x) = \int_{M_\Lambda} y\ d\rho(y|x) \tag{9}$$

satisfies

$$f_\rho \in \mathcal{G} = \{f \in \mathcal{F} :\ \ f = arg \min_{f\in\mathcal{F}} \mathcal{E}_\rho(f)\}, \tag{10}$$

where $\mathcal{F}$ is the set of all the Λ-valued measurable functions on X.

Since the learning problem we study is nontrivial, we assume $\mathcal{E}_\rho(f_\rho) > 0$. Also, the capacity of $\mathcal{H}$ is borrowed to measure the learning rates.

Let $(\mathcal{B}, d)$ be a metric space with metric d and $\varepsilon > 0$. The covering number $\mathcal{N}(\varepsilon,\ \mathcal{B}, d)$ is defined by

$$\begin{aligned} &\mathcal{N}(\varepsilon,\ \mathcal{B}, d) \\ = \quad &\min\{n :\ \text{there is a covering of } \mathcal{B} \text{ by } n \\ &\text{balls of radius} \le \varepsilon\}.\end{aligned}$$

Let $\mathcal{B}_R :=\{f \in \mathcal{B} :\quad \|f\|_d \le R\}$ be the closed ball of radius R. If there is $s > 0$ and a constant $c_s > 0$ only depends upon s such that,

$$\mathcal{N}(\eta,\ \mathcal{B}_R,\ \ d) \le c_s\Big(\frac{R}{\eta}\Big)^s, \qquad \forall\eta > 0, \tag{11}$$

then, we say $\mathcal{B}$ has logarithmic complexity exponent s.

Basing on above notions we give a learning rate for algorithm (7).

Theorem 1.1. Let $\mathcal{H} = \mathcal{H}_K$ be a Λ-valued RKHS with reproducing kernel kernel $K(x, y), f_{z,\lambda}$ be the solution of (7) and f_ρ satisfy (10). If $\mathcal{H}$ has logarithmic complexity $s \ge 0$, then, for $0 < \delta \le \frac{2}{e^{\frac{2+\sqrt[s]{4\times 3^s c_s}}{32}}}$ and $\lambda \le k^2\ D(f_\rho,\ \lambda)$, with confidence $1 - \delta$, holds

$$\begin{aligned}&\sqrt{\mathcal{E}_\rho(f_{z,\lambda})} - \sqrt{\mathcal{E}_\rho(f_\rho)} \\ &\le \frac{96\ k^2 M \log\frac{4}{\delta} \times \sqrt[4]{D(f_\rho,\ \ \lambda)}}{\lambda^{\ s+2}\!\!\sqrt{m}} \\ &+\frac{D(f_\rho, \lambda)}{\sqrt{\mathcal{E}_\rho(f_\rho)}}, \qquad (12)\end{aligned}$$

where

$$D(f_\rho,\ \lambda) = \inf_{f\in\mathcal{H}} \Big(\mathcal{E}_\rho(f) - \mathcal{E}_\rho(f_\rho) + \lambda\|f\|^2_{\mathcal{H}}\Big).$$

We now give some analysis on (12).

- By Lemma 2.1 we know

  $$\begin{aligned}\mathcal{E}_\rho(f) - \mathcal{E}_\rho(f_\rho) &= 2Var[f(\cdot) - f_\rho(\cdot)],\\ &\qquad f \in \mathcal{H}.\end{aligned}$$

  This estimates involve the approximation problem in the metric of variance (see [5, 20]). It has been proved by [20] that for the weighted Sobolev space whose reproducing kernel corresponding to the Wiener sheet measure, the approximation order may attain $n^{-1/2}$. So,the right side of (12) tends to 0 if $\lambda \to 0$ and $\lambda \sqrt[s+1]{m} \to +\infty$. Therefore, in this case,

  $$\sqrt{\mathcal{E}_\rho(f_{z,\lambda})} - \sqrt{\mathcal{E}_\rho(f_\rho)} \longrightarrow 0$$

  and thus

  $$\mathcal{E}_\rho(f_{z,\lambda}) - \mathcal{E}_\rho(f_\rho) \longrightarrow 0,$$

  which shows that

  $$Var[f(\cdot) - f_\rho(\cdot)] \longrightarrow 0.$$

  For some particular Λ-valued RKHSs the result can be strengthened. In fact,if f_ρ satisfies $E(f_\rho) = 0$ and $\mathcal{H}_K$ is an Λ-valued RKHS with reproducing kernel $K(x, y)$ and $\mathcal{H} = \{f \in \mathcal{H}_K : \int_X f(x)d\rho_X = 0\}$. Then, we have by the Lemma 2.6 in Section 2 that

  $$\begin{aligned}&\|f_{z,\lambda} - f_\rho\|_{L^2(\rho_X)}\\ \le\ & \|f_{z,\lambda} - f_{\rho,\lambda}\|_{L^2(\rho_X)} + \|f_{\rho,\lambda} - f_\rho\|_{L^2(\rho_X)}\\ \le\ & \sqrt{Var(f_{z,\lambda} - f_{\rho,\lambda})} + \sqrt{D(f_\rho, \lambda)}\\ \le\ & \frac{96\ k^2\ M\ \log\frac{4}{\delta} \times \sqrt[4]{D(f_\rho,\ \lambda)}}{\lambda\ \sqrt[s+2]{m}}\\ & +\sqrt{D(f_\rho, \lambda)}.\end{aligned}$$

- A key skill used in present paper is the parallelogram laws about RKHS norm, which was extended to q-uniform convex Banach spaces in [37, 38] and an inequality called the q-uniform convex inequality was established, combining which with the method used in [29] and [27] we can extend Theorem 1.1 to the case of vector valued RKBS spaces (see [39, 41]).

- The loss function used in this paper is the least square loss,the method and the results can be established for the other losses,for examples,the p-loss and the Lipschits loss (see [28, 30]).

2 Proofs

To show Theorem 1.1, we need some concepts and lemmas.

Let $(\mathcal{H}, \ \|\cdot\|_{\mathcal{H}})$ be a Hilbert space, $F(f) : \ \mathcal{H} \to R\bigcup\{\mp\infty\}$ 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

$$\lim_{t\to 0} \frac{F(f + tg) - F(f)}{t} = \langle g, \quad \xi\rangle_{\mathcal{H}} \qquad (13)$$

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

By Proposition 17.4 of [6] we know if $F(f)$: $\mathcal{H} \to R\bigcup\{\mp\infty\}$ is a convex function, then, $F(f)$ attains minimal value at f_0 if and only if $\nabla F(f_0) = 0$. Also, if $F(f)$: $\mathcal{H} \to R\bigcup\{\mp\infty\}$ is a Gâteaux differentiable function,then,by the Proposition 17.10 and Proposition 17.12 of [6] we know $F(f)$ is a convex on $\mathcal{H}$ if and only if for any $f, g \in \mathcal{H}$ we have

$$F(g + f) - F(f) \ge \Big\langle g,\ \nabla F(f)\Big\rangle_{\mathcal{H}}. \qquad (14)$$

(14) will be used to show the convexity of $\mathcal{E}_\rho(f)$.

Lemma 2.1. Let f_ρ be defined as in (9). Then, for any $f \in L^2(\rho_X)$, there holds

$$\mathcal{E}_\rho(f) - \mathcal{E}_\rho(f_\rho) = 2\ Var[f(\cdot) - f_\rho(\cdot)], \qquad (15)$$

where $Varf = \int_X (f(x) - E(f))^2 d\rho_X$ denotes the variance of f.

Proof. By the equality

$$\|a - b\|^2_\Lambda = \|a\|^2_\Lambda + \|b\|^2_\Lambda - 2\langle a,\ b\rangle_\Lambda$$

we have

$$\begin{aligned}\mathcal{E}_\rho(f) &= \int_Z\int_Z \|(y - f(x)) - (y' - f(x'))\|^2_\Lambda\\ &\quad \times d\rho(x, y)d\rho(x', y')\\ &= 2\Big(\int_Z\int_Z \|y - f(x)\|^2_\Lambda d\rho(x, y)\ d\rho(x', y')\\ &\quad - \int_Z\int_Z \langle y - f(x), y' - f(x')\rangle_\Lambda\\ &\quad \times d\rho(x, y)d\rho(x', y')\Big).\end{aligned}$$

Since

$$\begin{aligned}&\int_Z\int_Z \langle y - f(x), y' - f(x')\rangle_\Lambda\ d\rho(x, y)\ d\rho(x', y')\\ &= \|\int_Z (y - f(x))\ d\rho\|^2_\Lambda, \qquad (16)\end{aligned}$$

we have

$$\begin{aligned}\mathcal{E}_\rho(f) &= 2\Big(\int_Z \|y-f(x)\|_\Lambda^2 d\rho \\ &\quad -\|\int_Z (y-f(x))\, d\rho\|_\Lambda^2\Big) \\ &= 2\, Var[\,\cdot - f(\cdot)].\end{aligned}$$

By the definition of Var we have

$$\begin{aligned}& Var[\,\cdot - f(\cdot)] \\ = \;& Var\,[\cdot - f_\rho(\cdot) + f_\rho(\cdot) - f(\cdot)] \\ = \;& \int_Z \|(y - f_\rho(x) - E(\cdot - f_\rho(\cdot))) \\ + \;& (f_\rho(x) - f(x) - E(f_\rho(\cdot) - f(\cdot)))\|_\Lambda^2 \, d\rho \\ = \;& \int_Z \|y - f_\rho(x) - E(\cdot - f_\rho(\cdot))\|_\Lambda^2 \, d\rho \\ + \;& \int_Z \|f_\rho(x) - f(x) - E(f_\rho(\cdot) - f(\cdot))\|_\Lambda^2 d\rho \\ + \;& 2\int_Z \Big\langle y - f_\rho(x) - E(\cdot - f_\rho(\cdot)), \\ & f_\rho(x) - f(x) - E(f_\rho(\cdot) - f(\cdot))\Big\rangle_\Lambda \, d\rho.\end{aligned}$$

Since $E(\cdot - f_\rho(\cdot)) = 0$, we have by above formulas that

$$\begin{aligned}& Var[\,\cdot - f(\cdot)] \\ = \;& \int_Z \|y - f_\rho(x)\|_\Lambda^2 \, d\rho + \int_Z \|f_\rho(x) - f(x) \\ & -E(f_\rho(\cdot) - f(\cdot))\|_\Lambda^2 \, d\rho \\ & +2\int_Z \Big\langle y - f_\rho(x), \quad f_\rho(x) - f(x) \\ & -E(f_\rho(\cdot) - f(\cdot))\Big\rangle_\Lambda \, d\rho \\ = \;& \int_Z \|y - f_\rho(x)\|_\Lambda^2 d\rho + \int_Z \|f_\rho(x) - f(x) \\ & -E(f_\rho(\cdot) - f(\cdot))\|_\Lambda^2 \, d\rho \\ = \;& \int_Z \|y - f_\rho(x)\|_\Lambda^2 \, d\rho + Var\Big[f_\rho(\cdot) - f(\cdot)\Big],\end{aligned}$$

where we have used the fact

$$\begin{aligned}& \int_Z \Big\langle y - f_\rho(x), f_\rho(x) - f(x) - E(f_\rho(\cdot) - f(\cdot))\Big\rangle_\Lambda \\ & \times \, d\rho_X \\ & = \int_X \langle \int_Y (y - f_\rho(x)) \, d\rho(y|x), f_\rho(x) - f(x) \\ & -E(f_\rho(\cdot) - f(\cdot))\rangle_\Lambda \, d\rho_X = 0.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathcal{E}_\rho(f) &= 2\int_Z \|y - f_\rho(x)\|_\Lambda^2 \, d\rho \\ &\quad +2\, Var\,[f_\rho(\cdot) - f(\cdot)]. \qquad (17)\end{aligned}$$

Also, since

$$\begin{aligned}& \int_Z \int_Z \langle y - f_\rho(x), y' - f_\rho(x')\rangle_\Lambda d\rho(x,y) \, d\rho(x',y') \\ & \quad = 0,\end{aligned}$$

we have

$$\begin{aligned}\mathcal{E}_\rho(f_\rho) &= \int_Z \int_Z \Big\|(y - f_\rho(x)) - (y' - f_\rho(x'))\Big\|_\Lambda^2 \\ &\quad \times d\rho(x,y) \, d\rho(x',y') \\ &= \int_Z \int_Z \|y - f_\rho(x)\|_\Lambda^2 \, d\rho(x,y) \, d\rho(x',y') \\ &+ \int_Z \int_Z \|y' - f_\rho(x')\|_\Lambda^2 d\rho(x,y) \, d\rho(x',y') \\ &+ \int_Z \int_Z \langle y - f_\rho(x), \quad y' - f_\rho(x')\rangle_\Lambda \\ &\quad \times d\rho(x,y) \, d\rho(x',y') \\ &= 2\int_Z \|y - f_\rho(x)\|_\Lambda^2 \, d\rho. \qquad (18)\end{aligned}$$

(17) and (18) give (15).

Lemma 2.2. For $f \in \mathcal{H}$, we have

$$\begin{aligned}\nabla \mathcal{E}_\rho(f) &= -2 \int_Z \int_Z \Big(K(x,\cdot) - K(x',\cdot)\Big) \\ &\quad \times\Big[(y - y') - (f(x) - f(x'))\Big] \\ &\quad \times d\rho(x,y) \, d\rho(x',y'). \qquad (19)\end{aligned}$$

Proof. Since Λ is a Hilbert space, we have

$$\begin{aligned}\|a\|_\Lambda^2 - \|b\|_\Lambda^2 &= 2\langle a - b, \quad b\rangle_\Lambda + \|a - b\|_\Lambda^2, \\ &\qquad a, b \in \Lambda. \qquad (20)\end{aligned}$$

Then,

$$\begin{aligned}& \|[y - (f(x) + tg(x))] - [y' - (f(x') + tg(x'))]\|_\Lambda^2 \\ & -\|(y - f(x)) - (y' - f(x'))\|_\Lambda^2 \\ & = -2t\Big\langle g(x) - g(x'), (y - f(x)) - (y' - f(x'))\Big\rangle_\Lambda \\ & +t^2 \, \|g(x) - g(x')\|_\Lambda^2.\end{aligned}$$

Therefore, by the definition of Gatêaux derivative we have

$$\begin{aligned}& \lim_{t\to 0} \frac{\mathcal{E}_\rho(f + tg) - \mathcal{E}_\rho(f)}{t} \\ & = \lim_{t\to 0} \frac{1}{t} \int_Z \int_Z \Big(\|[y - (f(x) + tg(x))] \\ & -[y' - (f(x') + tg(x'))]\|_\Lambda^2 \\ & -\|(y - f(x)) - (y' - f(x'))\|_\Lambda^2\Big) \, d\rho(x,y) \\ & \times \, d\rho(x',y')\end{aligned}$$

$$= -2 \int_Z \int_Z \Big\langle g(x) - g(x'), \ (y - f(x)) \\ -(y' - f(x'))\Big\rangle_\Lambda d\rho(x,y) d\rho(x',y') \\ = -2 \int_Z \int_Z \Big\langle g(x), (y - f(x)) - (y' - f(x'))\Big\rangle_\Lambda \\ \times d\rho(x,y) \ d\rho(x',y') \\ +2 \int_Z \int_Z \Big\langle g(x'), (y - f(x)) - (y' - f(x'))\Big\rangle_\Lambda \\ \times \ d\rho(x,y) \ d\rho(x', \ y').$$

By (5) we know that above equality

$$= -2 \int_Z \int_Z \Big(g, K(x,\cdot)[(y - f(x)) \\ -(y' - f(x'))]\Big)_\mathcal{H} \ d\rho(x,y) \ d\rho(x',y') \\ +2 \int_Z \int_Z \Big(g, \ \ K(x',\cdot)[(y - f(x)) \\ -(y' - f(x'))]\Big)_\mathcal{H} \ d\rho(x,y) \ d\rho(x', \ y') \\ = \Big(g, -2 \int_Z \int_Z (K(x,\cdot) - K(x',\cdot)) \\ \times [(y - f(x)) - (y' - f(x'))] \\ \times \ d\rho(x,y) \ d\rho(x',y')\Big)_\mathcal{H}. \tag{21}$$

(21) gives (19).

We show further the following result:

Lemma 2.3. $\mathcal{E}_\rho(f)$ is a convex function on $\mathcal{H}$.

Proof. By (20) we have

$$\|a\|_\Lambda^2 - \|b\|_\Lambda^2 \geq \langle a - b, \ 2b\rangle_\Lambda, \qquad a, \ b \in \Lambda. \tag{22}$$

Therefore, for any $f, \ g \in \mathcal{H}$ we have

$$\|(y - g(x)) - (y' - g(x'))\|_\Lambda^2 \\ -\|(y - f(x)) - (y' - f(x'))\|_\Lambda^2 \\ \geq \ \Big\langle (g(x) - f(x)) - (g(x') - f(x')), \\ -2[(y - f(x)) - (y' - f(x'))]\Big\rangle_\Lambda$$

and

$$\mathcal{E}_\rho(g) - \mathcal{E}_\rho(f) \\ \geq \ \int_Z \int_Z \Big\langle (g(x) - f(x)) - (g(x') - f(x')), \\ -2[(y - f(x)) - (y' - f(x'))]\Big\rangle_\Lambda \\ \times \ d\rho(x,y) \ d\rho(x',y') \\ = \ \int_Z \int_Z \Big\langle g(x) - f(x), \ \ -2[(y - f(x)) \\ -(y' - f(x'))]\Big\rangle_\Lambda \ d\rho(x,y) \ d\rho(x',y') \\ - \int_Z \int_Z \Big\langle (g(x') - f(x'), \ \ -2[(y - f(x)) \\ -(y' - f(x'))]\Big\rangle_\Lambda \ d\rho(x,y) \ d\rho(x',y')$$

Also, by (5) we know above equality

$$= \ \int_Z \int_Z \Big(g - f, -2K(\cdot,x)[(y - f(x)) \\ -(y' - f(x'))]\Big)_\mathcal{H} d\rho(x,y) \ d\rho(x',y') \\ - \ \int_Z \int_Z \Big(g - f, \ \ -2K(\cdot, \ x') \\ \times [(y - f(x)) - (y' - f(x'))]\Big)_\mathcal{H} \\ \times d\rho(x,y) \ d\rho(x',y') \\ = \ \Big(g - f, -2 \int_Z \int_Z K(\cdot, \ x)[(y - f(x)) \\ -(y' - f(x'))]d\rho(x,y) \ d\rho(x',y')\Big)_\mathcal{H} \\ -\Big(g - f, -2 \int_Z \int_Z K(\cdot, \ x')[(y - f(x)) \\ -(y' - f(x'))]d\rho(x,y) \ d\rho(x',y')\Big)_\mathcal{H} \\ = \ \Big(g - f, -2 \int_Z \int_Z (K(\cdot, \ x) - K(\cdot, \ x')) \\ \times [(y - f(x)) - (y' - f(x'))]d\rho(x,y) \\ \times d\rho(x',y')\Big)_\mathcal{H} \\ = \ \Big(g - f, \nabla \mathcal{E}_\rho(f)\Big)_\mathcal{H}. \tag{23}$$

By (14) and (23) we know $\mathcal{E}_\rho(f)$ is a convex function on $\mathcal{H}$. But $\mathcal{E}_\rho(f)$ is not a strict convex since $\mathcal{G}$ is not a single set.

Define the integral form of (7) by

$$f_{\rho,\lambda} = arg \min_{f \in \mathcal{H}} \mathcal{E}_\rho(f) + \lambda \|f\|_\mathcal{H}^2. \tag{24}$$

Since $\|f\|_\mathcal{H}^2$ is strict convex about f on $\mathcal{H}$, we know $\mathcal{E}_\rho(f) + \lambda \ \|f\|_\mathcal{H}^2$ is a strict convex function on $\mathcal{H}$, the solution $f_{\rho,\lambda}$ is unique.

Lemma 2.4. Let $f_{z,\lambda}$ be the solution of (7) and $f_{\rho,\lambda}$ be the solution of (24). Then, there hold inequalities

$$\|f_{z,\lambda}\|_\mathcal{H} \leq M\sqrt{\frac{2}{\lambda}}. \tag{25}$$

and

$$\|f_{\rho,\lambda}\|_\mathcal{H} \leq \sqrt{\frac{D(f_\rho, \lambda)}{\lambda}}. \tag{26}$$

Also, $f_{\rho, \ \lambda}$ satisfies the equality

$$\lambda f_{\rho, \ \lambda}(\cdot) \ = \ \int_Z \int_Z \Big(K(x,\cdot) - K(x',\cdot)\Big) \\ \times \ \ [(y - y') - (f_{\rho,\lambda}(x) - f_{\rho,\lambda}(x'))] \\ \times \ d\rho(x,y) \ d\rho(x',y'). \tag{27}$$

Proof.Proof of (25). By the definition of $f_{\rho,\lambda}$ and f_ρ we have

$$\begin{aligned} &\mathcal{E}_z(f_{z,\lambda}) + \lambda\|f_{z,\lambda}\|^2_{\mathcal{H}} \\ \leq\ & \mathcal{E}_z(0) \\ =\ & \frac{1}{m(m-1)}\sum_{i,j=1,i\neq j}^{m} \|y_i-y_j\|^2_\Lambda \leq 2M^2, \end{aligned}$$

which gives (25).

Proof of (26). By the definition of f_ρ we have

$$\mathcal{E}_\rho(f_{\rho,\lambda}) - \mathcal{E}_\rho(f_\rho) \geq 0.$$

Hence,

$$\begin{aligned}\lambda\|f_{\rho,\lambda}\|^2_{\mathcal{H}} &\leq \mathcal{E}_\rho(f_{\rho,\lambda}) - \mathcal{E}_\rho(f_\rho)+\lambda\|f_{\rho,\lambda}\|^2_{\mathcal H}\\ &= D(f_\rho,\ \lambda).\end{aligned}$$

(26) then holds.

Proof of (27). By the definition of $f_{\rho,\lambda}$ we have

$$\begin{aligned} 0 &= \nabla\Big(\mathcal{E}_\rho(f)+\lambda\|f\|^2_{\mathcal H}\Big)|_{f=f_{\rho,\lambda}}\\ &= \nabla\mathcal{E}_\rho(f)|_{f=f_{\rho,\lambda}} + \lambda\nabla(\|f\|^2_{\mathcal H})|_{f=f_{\rho,\lambda}}\\ &= -2\int_Z\int_Z\Big(K(x,\cdot)-K(x',\cdot)\Big)\\ &\quad\times\Big[(y-y')-(f_{\rho,\lambda}(x)-f_{\rho,\lambda}(x'))\Big]\\ &\quad\times\, d\rho(x,y)\; d\rho(x',y')\\ &\quad+2\ \lambda\ f_{\rho,\lambda}(\cdot). \end{aligned}$$

(27) thus holds.

Following Lemma 2.5 quantitatively shows the dependence of $f_{\rho,\ \lambda}$ upon the distributions ρ.

Lemma 2.5. Let $f_{\rho,\lambda}$ and $f_{\gamma,\lambda}$ be the solutions of (24) w.r.t.(with respect to) the distributions ρ and γ respectively. Then,

$$\begin{aligned} &\lambda\|f_{\rho,\lambda}-f_{\gamma,\lambda}\|_{\mathcal H} + \frac{2Var(f_{\rho,\lambda}-f_{\gamma,\lambda})}{\|f_{\rho,\lambda}-f_{\gamma,\lambda}\|_{\mathcal H}}\\ \leq\ & 2\Big\|\int_Z\int_Z[K(x,\ \cdot)-K(x',\cdot)]\\ &\times[(y-y')-(f_{\rho,\lambda}(x)-f_{\rho,\lambda}(x'))]\ d\rho(x,y)\\ &\times d\rho(x',y')\\ &-\int_Z\int_Z[K(x,\cdot)-K(x',\cdot)]\\ &\times[(y-y')-(f_{\rho,\lambda}(x)-f_{\rho,\lambda}(x'))]\\ &\times d\gamma(x,y)d\gamma(x',y')\Big\|_{\mathcal H}. \qquad (28) \end{aligned}$$

Proof. By (23) and the equality (20) we have

$$\begin{aligned} &\mathcal{E}_\gamma(f_{\gamma,\lambda})-\mathcal{E}_\gamma(f_{\rho,\lambda})\\ =\ & \Big(f_{\gamma,\lambda}-f_{\rho,\lambda},-2\int_Z\int_Z(K(x,\cdot)-K(x',\cdot))\\ \times\ & [(y-y')-(f_{\rho,\lambda}(x)-f_{\rho,\lambda}(x'))]\\ &\times d\rho(x,y)\ d\rho(x',y')\Big)_{\mathcal H}\\ &+2Var(f_{\rho,\lambda}-f_{\gamma,\lambda}). \qquad (29) \end{aligned}$$

Therefore, by the equality (20) of Hilbert space $\mathcal H$ we have

$$\begin{aligned} 0\ \geq\ & (\mathcal{E}_\gamma(f_{\gamma,\lambda})+\lambda\|f_{\gamma,\lambda}\|^2_{\mathcal H})-(\mathcal{E}_\gamma(f_{\rho,\lambda})+\lambda\|f_{\rho,\lambda}\|^2_{\mathcal H})\\ =\ & (f_{\gamma,\lambda}-f_{\rho,\lambda},-2\int_Z\int_Z(K(x,\cdot)-K(x',\cdot))\\ &\times[(y-y')-(f_{\rho,\lambda}(x)-f_{\rho,\lambda}(x'))]\\ &\times\ d\gamma(x,y)\ d\gamma(x',y'))_{\mathcal H}\\ &+(f_{\gamma,\lambda}-f_{\rho,\lambda},\ 2\lambda f_{\rho,\lambda})_{\mathcal H}+\lambda\ \|f_{\gamma,\lambda}-f_{\rho,\lambda}\|^2_{\mathcal H}\\ &+2Var(f_{\rho,\lambda}-f_{\gamma,\lambda}). \end{aligned}$$

By (27) we have

$$\begin{aligned} 0\ \geq\ & \Big(f_{\gamma,\lambda}-f_{\rho,\lambda},\ -2\int_Z\int_Z(K(x,\cdot)-K(x',\cdot))\\ \times\ & [(y-y')-(f_{\rho,\lambda}(x)-f_{\rho,\lambda}(x'))]\\ &\times d\gamma(x,y)\ d\gamma(x',y')\Big)_{\mathcal H}\\ &+\Big(f_{\gamma,\lambda}-f_{\rho,\lambda},2\int_Z\int_Z(K(x,\cdot)-K(x',\cdot))\\ \times\ & [(y-y')-(f_{\rho,\lambda}(x)-f_{\rho,\lambda}(x'))]\\ &\times\ d\rho(x,y)\ d\rho(x',y')\Big)_{\mathcal H}+\lambda\|f_{\gamma,\lambda}-f_{\rho,\lambda}\|^2_{\mathcal H}\\ &+2Var(f_{\rho,\lambda}-f_{\gamma,\lambda})\\ =\ & 2\Big(f_{\gamma,\lambda}-f_{\rho,\lambda},\int_Z\int_Z(K(x,\cdot)-K(x',\cdot))\\ &\times[(y-y')-(f_{\rho,\lambda}(x)-f_{\rho,\lambda}(x'))]\ d\rho(x,y)\\ &\times d\rho(x',y')\\ &-\int_Z\int_Z(K(x,\cdot)-K(x',\cdot))\\ &\times[(y-y')-(f_{\rho,\lambda}(x)-f_{\rho,\lambda}(x'))]\\ &\times d\gamma(x,y)\ d\gamma(x',y')\Big)_{\mathcal H}\\ &+\lambda\ \|f_{\gamma,\lambda}-f_{\rho,\lambda}\|^2_{\mathcal H}+2Var(f_{\rho,\lambda}-f_{\gamma,\lambda}). \end{aligned}$$

It follows by the Cauchy's inequality that

$$\begin{aligned} &\lambda\|f_{\gamma,\lambda}-f_{\rho,\lambda}\|^2_{\mathcal H}+2Var(f_{\rho,\lambda}-f_{\gamma,\lambda})\\ &\leq 2\Big(f_{\rho,\lambda}-f_{\gamma,\lambda},\int_Z\int_Z(K(x,\cdot)-K(x',\cdot))\\ &\times[(y-y')-(f_{\rho,\lambda}(x)-f_{\rho,\lambda}(x'))]\ d\rho(x,y)\\ &\times d\rho(x',\ y')\\ &-\int_Z\int_Z(K(x,\cdot)-K(x',\cdot))\\ &\times[(y-y')-(f_{\rho,\lambda}(x)-f_{\rho,\lambda}(x'))]\ d\gamma(x,y) \end{aligned}$$

$$\begin{aligned}
&\times d\gamma(x',y')\Big)_{\mathcal{H}}\\
&\le 2\|f_{\rho,\lambda}-f_{\gamma,\lambda}\|_{\mathcal{H}}\\
&\times\Big\|\int_Z\int_Z (K(x,\cdot)-K(x',\cdot))\\
&\times[(y-y')-(f_{\rho,\lambda}(x)-f_{\rho,\lambda}(x'))]\,d\rho(x,y)\\
&\times d\rho(x',y')\\
&-\int_Z\int_Z (K(x,\cdot)-K(x',\cdot))\\
&\times[(y-y')-(f_{\rho,\lambda}(x)-f_{\rho,\lambda}(x'))]\,d\gamma(x,y)\\
&\times d\gamma(x',y')\Big\|_{\mathcal{H}}. \qquad (30)
\end{aligned}$$

(30) gives (28).

Lemma 2.6. Let $f_{z,\lambda}$ be the solution of (7) and $f_{\rho,\lambda}$ be the solution of (24). If $(\mathcal{H},\ \|\cdot\|_{\mathcal{H}})$ has logarithmic complexity $s\ge 0$, then, for any $0<\delta\le \frac{2}{e^{\frac{\sqrt[s+2]{4\times 3^s\ c_s}}{32}}}$ and $\lambda\le k^2\ \ D(f_\rho,\ \lambda)$, with confidence $1-\delta$, holds

$$\begin{aligned}
&\sqrt{Var(f_{z,\lambda}-f_{\rho,\lambda})}\\
\le\ & \frac{48\,k^2\ M\ \log\frac{4}{\delta}\times\sqrt[4]{D(f_\rho,\lambda)}}{\lambda\ \sqrt[s+2]{m}}. \qquad (31)
\end{aligned}$$

Proof. Define an empirical distribution γ_z as

$$\begin{aligned}
&\int_Z\int_Z f\Big[(x,y),\ (x',\ y')\Big]\,d\gamma_z\\
&=\frac{1}{m(m-1)}\sum_{i,j=1,i\ne j}^{m} f\Big[(x_i,\ y_i),\ (x_j,y_j)\Big].
\end{aligned}$$

Then, by (28) we have

$$\begin{aligned}
&\|f_{z,\lambda}-f_{\rho,\lambda}\|_{\mathcal{H}}\\
\le\ & \frac{2}{\lambda}\Big\|\int_Z\int_Z (K(x,\cdot)-K(x',\cdot))\\
&\times[(y-y')-(f_{z,\lambda}(x)-f_{z,\lambda}(x'))]\,d\rho(x,y)\\
&\times\ d\rho(x',y')\\
-\ & \frac{1}{m(m-1)}\sum_{i,j=1,i\ne j}^{m}(K(x_i,\cdot)-K(x_j,\cdot))\\
&\times[(y_i-y_j)-(f_{z,\lambda}(x_i)-f_{z,\lambda}(x_j))]\Big\|_{\mathcal{H}}
\end{aligned}$$

and

$$\begin{aligned}
&\frac{Var(f_{\rho,\lambda}-f_{z,\lambda})}{\|f_{\rho,\lambda}-f_{z,\lambda}\|_{\mathcal{H}}}\\
\le\ & \Big\|\int_Z\int_Z (K(x,\cdot)-K(x',\cdot))\\
&\times[(y-y')-(f_{z,\lambda}(x)-f_{z,\lambda}(x'))]\,d\rho(x,y)\\
&\times d\rho(x',y')\\
-\ & \frac{1}{m(m-1)}\sum_{i,j=1,i\ne j}^{m}(K(x_i,\cdot)-K(x_j,\cdot))\\
&\times[(y_i-y_j)-(f_{z,\lambda}(x_i)-f_{z,\lambda}(x_j))]\Big\|_{\mathcal{H}}.
\end{aligned}$$

Above two inequalities give

$$\begin{aligned}
&Var(f_{z,\lambda}-f_{\rho,\lambda})\\
&=\frac{Var(f_{z,\lambda}-f_{\rho,\lambda})}{\|f_{z,\lambda}-f_{\rho,\lambda}\|_{\mathcal{H}}}\times\|f_{z,\lambda}-f_{\rho,\lambda}\|_{\mathcal{H}}\\
&\le\frac{2}{\lambda}\Big\|\int_Z\int_Z (K(x,\cdot)-K(x',\cdot))\\
&\times[(y-y')-(f_{z,\lambda}(x)-f_{z,\lambda}(x'))]\,d\rho(x,y)\,d\rho(x',y')\\
&-\frac{1}{m(m-1)}\sum_{i,j=1,i\ne j}^{m}(K(x_i,\cdot)-K(x_j,\cdot))\\
&\times[(y_i-y_j)-(f_{z,\lambda}(x_i)-f_{z,\lambda}(x_j))]\Big\|_{\mathcal{H}}\\
&\times\Big\|\int_Z\int_Z (K(x,\cdot)-K(x',\cdot))\\
&\times[(y-y')-(f_{\rho,\lambda}(x)-f_{\rho,\lambda}(x'))]\,d\rho(x,y)\,d\rho(x',y')\\
&-\frac{1}{m(m-1)}\sum_{i,j=1,i\ne j}^{m}(K(x_i,\cdot)-K(x_j,\cdot))\\
&\times[(y_i-y_j)-(f_{\rho,\lambda}(x_i)-f_{\rho,\lambda}(x_j))]\Big\|_{\mathcal{H}}\\
&=\frac{2}{\lambda}A(z)\times B(z), \qquad (32)
\end{aligned}$$

where

$$\begin{aligned}
A(z)\ =\ & \Big\|\int_Z\int_Z (K(x,\cdot)-K(x',\cdot))\\
&\times[(y-y')-(f_{z,\lambda}(x)-f_{z,\lambda}(x'))]\,d\rho(x,y)\\
&\times\ d\rho(x',y')\\
&-\frac{1}{m(m-1)}\sum_{i,j=1,i\ne j}^{m}(K(x_i,\cdot)-K(x_j,\cdot))\\
&\times[(y_i-y_j)-(f_{z,\lambda}(x_i)-f_{z,\lambda}(x_j))]\Big\|_{\mathcal{H}},
\end{aligned}$$

and

$$\begin{aligned}
B(z)\ =\ & \Big\|\int_Z\int_Z (K(x,\cdot)-K(x',\cdot))\\
&\times[(y-y')-(f_{\rho,\lambda}(x)-f_{\rho,\lambda}(x'))]\,d\rho(x,y)\\
&\times d\rho(x',y')\\
&-\frac{1}{m(m-1)}\sum_{i,j=1,i\ne j}^{m}(K(x_i,\cdot)-K(x_j,\cdot))\\
&\times[(y_i-y_j)-(f_{\rho,\lambda}(x_i)-f_{\rho,\lambda}(x_j))]\Big\|_{\mathcal{H}}.
\end{aligned}$$

Since $\|f\|_{\mathcal{H}}=\sup\limits_{\|h\|_{\mathcal{H}}\le 1}(h,f)_H$, we have

$$A(z)\ =\ \sup_{\|h\|_{\mathcal{H}}\le 1}\Big|\Big(h,\int_Z\int_Z (K(x,\cdot)-K(x',\cdot))$$

$$\begin{aligned}
&\times[(y-y')-(f_{z,\lambda}(x)-f_{z,\lambda}(x'))]\,d\rho(x,y)\\
&\times\,d\rho(x',y')\\
&-\frac{1}{m(m-1)}\sum_{i,j=1,i\neq j}^{m}(K(x_i,\cdot)-K(x_j,\cdot))\\
&\times[(y_i-y_j)-(f_{z,\lambda}(x_i)-f_{z,\lambda}(x_j))]\Big)_{\mathcal{H}}\Big|\\
=\;&\sup_{\|h\|_{\mathcal{H}}\leq 1}\Big|\int_Z\int_Z\Big(h,\;(K(x,\cdot)-K(x',\cdot))\\
&\times[(y-y')-(f_{z,\lambda}(x)-f_{z,\lambda}(x'))]\Big)_{\mathcal{H}}\\
&\quad\times\,d\rho(x,y)\,d\rho(x',y')\\
&-\frac{1}{m(m-1)}\sum_{i,j=1,i\neq j}^{m}\Big(h,(K(x_i,\cdot)-K(x_j,\cdot))\\
&\times[(y_i-y_j)-(f_{z,\lambda}(x_i)-f_{z,\lambda}(x_j))]\Big)_{\mathcal{H}}\Big|\\
\overset{(5)}{=}\;&\sup_{\|h\|_{\mathcal{H}}\leq 1}\Big|\int_Z\int_Z\Big\langle h(x)-h(x'),\\
&[(y-y')-(f_{z,\lambda}(x)-f_{z,\lambda}(x'))]\Big\rangle_{\Lambda}\\
&\times d\rho(x,y)\,d\rho(x',\,y')\\
&-\frac{1}{m(m-1)}\sum_{i,j=1,i\neq j}^{m}\Big\langle h(x_i)-h(x_j),\\
&\;[(y_i-y_j)-(f_{z,\lambda}(x_i)-f_{z,\lambda}(x_j))]\Big\rangle_{\Lambda}\Big|.
\end{aligned}$$

Also, since $\|h\|_{\mathcal{H}}\leq 1$ we have by (25) and (6) that

$$\begin{aligned}
&\Big|\xi[(x,y),(x',y'),h]\Big|\\
=\;&\Big|\Big\langle h(x)-h(x'),(y-y')\\
&-(f_{z,\lambda}(x)-f_{z,\lambda}(x'))\Big\rangle_{\Lambda}\Big|\\
\leq\;&\Big\|h(x)-h(x')\Big\|_{\Lambda}\\
&\times\Big\|(y-y')-(f_{z,\lambda}(x)-f_{z,\lambda}(x'))\Big\|_{\Lambda}\\
\leq\;&(\|h(x)\|_{\Lambda}+\|h(x')\|_{\Lambda})\\
&\times(2M+\|f_{z,\lambda}(x)\|_{\Lambda}+\|f_{z,\lambda}(x')\|_{\Lambda})\\
\leq\;&4k\|h\|_{\mathcal{H}}\Big(M+kM\sqrt{\frac{2}{\lambda}}\Big)\\
\leq\;&4kM\Big(1+k\sqrt{\frac{2}{\lambda}}\Big)
\end{aligned}$$

and

$$\begin{aligned}
&\Big\|\int_Z\int_Z(K(x,\cdot)-K(x',\cdot))\\
&\times[(y-y')-(f_{z,\lambda}(x)-f_{z,\lambda}(x'))]\\
&\times\,d\rho(x,y)\,d\rho(x',y')\\
&-\frac{1}{m(m-1)}\sum_{i,j=1,i\neq j}^{m}(K(x_i,\cdot)-K(x_j,\cdot))\\
&\times[(y_i-y_j)-(f_{z,\lambda}(x_i)-f_{z,\lambda}(x_j))]\Big\|_{\mathcal{H}}\\
&\leq\sup_{\xi\in\mathcal{B}}\Big|\int_Z\int_Z\xi[(x,y),(x',y')]\,d\rho(x,y)\,d\rho(x',y')\\
&-\frac{1}{m(m-1)}\sum_{i,j=1,i\neq j}^{m}\xi[(x_i,y_i),(x_j,y_j)]\Big|,
\end{aligned}$$

where

$$\begin{aligned}
\mathcal{B}\;=\;&\Big\{\xi[(x,y),(x',y'),h]\\
:\;&\Big|\xi[(x,y),\;(x',y'),\;h]\Big|\leq 4kM\Big(1+k\sqrt{\frac{2}{\lambda}}\Big)\\
&\text{for any }[(x,y),\;(x',\;y'),h]\in Z\times Z\times\mathcal{H}\Big\}.
\end{aligned}$$

Let $N=\mathcal{N}(\mathcal{B},\quad\varepsilon)$ be the covering number of $\mathcal{B}$ for $\varepsilon>0$. Then, by the formal method used in learning theory, there are $\{\xi_j\}_{j=1}^{N}\subset\mathcal{B}$ such that for any $\xi\in\mathcal{B}$ we have a $\xi_k\in\{\xi_j\}_{j=1}^{N}$ such that $\|\xi-\xi_k\|_{C(X,\Lambda)}<\varepsilon$

$$\begin{aligned}
&Prob\Big\{\sup_{\xi\in\mathcal{B}}\Big|\int_Z\int_Z\xi[(x,y),\quad(x',y')]\,d\rho(x,y)\\
&\times d\rho(x',\;y')\\
&-\frac{1}{m(m-1)}\sum_{i,j=1,i\neq j}^{m}\xi[(x_i,y_i),(x_j,y_j)]\Big|>3\varepsilon\Big\}\\
&\leq\sum_{k=1}^{N}Prob\Big\{\Big|\int_Z\int_Z\xi_k[(x,y),(x',y')]\,d\rho(x,y)\\
&\times d\rho(x',y')\\
&-\frac{1}{m(m-1)}\sum_{i,j=1,i\neq j}^{m}\xi_k[(x_i,y_i),(x_j,y_j)]\Big|>\varepsilon\Big\}.
\end{aligned}$$

Recall the U-statistics inequality (see [3, 19]):

$$\begin{aligned}
&Prob\Big\{\Big|\frac{1}{m(m-1)}\sum_{i,j=1,i\neq j}^{m}U(z_i,\;z_j)\\
&-E(U)\Big|\geq\varepsilon\Big\}\\
&\leq 2\exp\Big\{-\frac{(m-1)\varepsilon^2}{4\sigma^2+(4/3)(b-a)\varepsilon}\Big\},
\end{aligned}$$

where $a\leq U(z,\quad z')\leq b$ and $Var[U]=\sigma^2$, by which we have

$$\begin{aligned}
&Prob\Big\{\sup_{\xi\in\mathcal{B}}\Big|\int_Z\int_Z\xi[(x,y),\quad(x',y')]\,d\rho(x,y)\\
&\times d\rho(x',y')\\
&-\frac{1}{m(m-1)}\sum_{i,j=1,i\neq j}^{m}\xi[(x_i,y_i),(x_j,y_j)]\Big|>3\varepsilon\Big\}\\
&\leq 2\mathcal{N}(\mathcal{B},\;\varepsilon)\exp\Big(-\frac{(m-1)\varepsilon^2}{4\tau^2+8\tau\varepsilon/3}\Big)\qquad(33)
\end{aligned}$$

Since $\mathcal{N}(\mathcal{H}_1,\ \varepsilon) \leq \exp\{c_s(\frac{1}{\varepsilon})^s\}$, we have by defining $\tau = 4kM\Big(1 + k\sqrt{\frac{2}{\lambda}}\Big)$ that

$$\log \mathcal{N}(\mathcal{B},\ \varepsilon) \ \ \leq \ \ \log \mathcal{N}(\mathcal{H}_1,\ \frac{\varepsilon}{\tau}) \leq c_s(\frac{\tau}{\varepsilon})^s,$$

i.e.,

$$\mathcal{N}(\mathcal{B},\ \varepsilon) \ \ \leq \ \ \exp\big(c_s(\frac{\tau}{\varepsilon})^s\big).$$

By (33) we have

$$\begin{aligned} & Prob\Big\{ \sup_{\xi \in \mathcal{B}} \Big| \int_Z \int_Z \xi[(x,y),(x',y')]\ d\rho(x,y) \\ & \times d\rho(x',y') \\ & -\frac{1}{m(m-1)} \sum_{i,j=1, i\neq j}^{m} \xi[(x_i,y_i),(x_j,y_j)] \Big| > 3\varepsilon \Big\} \\ & \leq 2\exp\Big\{ c_s(\frac{\tau}{\varepsilon})^s - \frac{(m-1)\varepsilon^2}{4\tau^2 + 8\tau\varepsilon/3} \Big\}, \end{aligned}$$

i.e.,

$$\begin{aligned} & Prob\Big\{ \sup_{\xi \in \mathcal{B}} \Big| \int_Z \int_Z \xi[(x,y),\ (x',y')]\ d\rho(x,y) \\ & \times d\rho(x',y') \\ & -\frac{1}{m(m-1)} \sum_{i,j=1, i\neq j}^{m} \xi[(x_i,y_i),(x_j,y_j)] \Big| > h \Big\} \\ & \leq 2\exp\Big\{ c_s(\frac{3\tau}{h})^s - \frac{(m-1)h^2}{36\tau^2 + 8\tau h} \Big\}. \end{aligned}$$

Take

$$2\exp\Big\{ c_s\ (\frac{3\tau}{h})^s - \frac{(m-1)h^2}{36\tau^2 + 8\tau h} \Big\} = \frac{\delta}{2}.$$

We have by simple computations that

$$\begin{aligned} & h^{2+s} - \frac{8\tau \log\frac{4}{\delta}}{m-1} h^{s+1} - \frac{36\tau^2 \log\frac{4}{\delta}}{m-1} h^s \\ & -\frac{8 \times 3^s c_s \tau^{s+1}}{m-1} h - \frac{3^s \times c_s \tau^{2+s}}{m-1} = 0. \quad (34) \end{aligned}$$

By Lemma 7.2 of [16] we have following result:

Let $c_1,\ c_2,\ \cdots,\ c_l > 0$ *and* $s > q_1 > q_2 > \cdots > q_{l-1} > 0$. *Then, the equation*

$$x^s - c_1 x^{q_1} - c_2 x^{q_2} - \cdots - c_{l-1} x^{q_{l-1}} - c_l = 0$$

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

$$\begin{aligned} x^* \ \ \leq \ \ & \max\Big\{ (lc_1)^{1/(s-q_1)}, (lc_2)^{1/(s-q_2)}, \cdots, \\ & (lc_{l-1})^{1/(s-q_{l-1})}, (lc_l)^{1/s} \Big\}. \qquad (35) \end{aligned}$$

Above inequality and (34) give

$$\begin{aligned} h \ \ & \leq \ \ h^* \\ & \leq \ \ \tau \times \max\Big(\frac{32\ \log\frac{4}{\delta}}{m-1}, 12\sqrt{\frac{\log\frac{4}{\delta}}{m-1}}, \\ & \qquad \sqrt[s+1]{\frac{32 \times 3^s c_s}{m-1}},\ \ \sqrt[s+2]{\frac{4 \times 3^s c_s}{m-1}} \Big) \\ & \leq \ \ \tau \times \max\Big(\frac{32 \log\frac{4}{\delta}}{\sqrt{m-1}}, 2\sqrt[s+2]{\frac{4 \times 3^s c_s}{m-1}} \Big) \\ & \leq \ \ \frac{3\tau \times \log\frac{4}{\delta}}{\sqrt[s+2]{m}} \\ & \leq \ \ \frac{12k\ M\ \log\frac{4}{\delta}}{\sqrt[s+2]{m}} \times \Big(1 + k\sqrt{\frac{2}{\lambda}}\Big) \\ & \leq \ \ \frac{12k^2\ M\ \log\frac{4}{\delta}}{\sqrt[s+2]{m}} \times \sqrt{\frac{2}{\lambda}} \end{aligned}$$

if $0 \ < \ \delta \ \leq \ \frac{2}{e^{\frac{2+\sqrt[s]{576 \times 6^s\ c_s}}{64}}}$ and $\lambda \ \leq \ k^2 D(f_\rho,\ \ \lambda)$. Therefore, with confidence $1 - \frac{\delta}{2}$, holds

$$A(z) \leq \frac{12k^2\ M \log\frac{4}{\delta}}{\sqrt[s+2]{m}} \times \sqrt{\frac{2}{\lambda}}. \qquad (36)$$

Repeating above procedure and use the fact (26) we have

$$B(z) \leq \frac{12k^2 M\ \log\frac{4}{\delta}}{\sqrt[s+2]{m}} \times \sqrt{\frac{D(f_\rho,\ \lambda)}{\lambda}}. \qquad (37)$$

By (32), (36) and (37) we have (31).

Proof of Theorem 1.1. By Minkowski inequality we have

$$\begin{aligned} & \Big| \sqrt{\mathcal{E}_\rho(f_{z,\lambda})} - \sqrt{\mathcal{E}_\rho(f_{\rho,\ \lambda})} \Big| \\ = \ \ & \Big| \Big(\int_Z \int_Z \Big\| (y - f_{z,\lambda}(x)) - (y' - f_{z,\ \lambda}(x')) \Big\|_\Lambda^2 \\ & \times d\rho(x,y)\ d\rho(x',y') \Big)^{\frac{1}{2}} \\ & -\Big(\int_Z \int_Z \Big\| (y - f_{\rho,\lambda}(x)) - (y' - f_{\rho,\lambda}(x')) \Big\|_\Lambda^2 \\ & \times d\rho(x,y)\ d\rho(x',y') \Big)^{\frac{1}{2}} \Big| \\ \leq \ \ & \Big(\int_Z \int_Z \Big\| (f_{\rho,\lambda}(x) - f_{z,\lambda}(x)) \\ & -(f_{\rho,\lambda}(x') - f_{z,\lambda}(x')) \Big\|_\Lambda^2\ d\rho(x,y)\ d\rho(x',y') \Big)^{\frac{1}{2}} \\ = \ \ & \sqrt{2Var(f_{\rho,\lambda} - f_{z,\lambda})}. \qquad (38) \end{aligned}$$

Also, since $\mathcal{E}_\rho(f_{z,\lambda}) \geq \mathcal{E}_\rho(f_\rho)$ and

$$\sqrt{b} - \sqrt{a} \leq \frac{b-a}{\sqrt{a}} \quad \text{for} \quad b > a > 0,$$

we have by (38) and (31) that

$$\begin{aligned}
& \sqrt{\mathcal{E}_\rho(f_{z,\lambda})} - \sqrt{\mathcal{E}_\rho(f_\rho)} \\
= \ & |\sqrt{\mathcal{E}_\rho(f_{z,\,\lambda})} - \sqrt{\mathcal{E}_\rho(f_\rho)}| \\
\le \ & |\sqrt{\mathcal{E}_\rho(f_{z,\lambda})} - \sqrt{\mathcal{E}_\rho(f_{\rho,\lambda})}| \\
& + \sqrt{\mathcal{E}_\rho(f_{\rho,\lambda})} - \sqrt{\mathcal{E}_\rho(f_\rho)} \\
\le \ & \sqrt{2Var(f_{z,\,\lambda} - f_{\rho,\lambda})} + \frac{\mathcal{E}_\rho(f_{\rho,\lambda}) - \mathcal{E}_\rho(f_\rho)}{\sqrt{\mathcal{E}_\rho(f_\rho)}} \\
\le \ & \frac{96\,k^2\,M\ \log\frac{4}{\delta} \times \sqrt[4]{D(f_\rho,\ \lambda)}}{\lambda\ \sqrt[s+2]{m}} \\
& + \frac{D(f_\rho,\ \lambda)}{\sqrt{\mathcal{E}_\rho(f_\rho)}}.
\end{aligned}$$

(12) is proved.

Acknowledgements: This research is supported by the National Natural Science Foundation of China under Grant no. 11471292,61179041 and 11101220.

References:

[1] S. Agarwal, D. Dugar,S. Sengupt, Ranking chemical structures for drug discovery:a new machine learning approach,*J. Chem. Inf. Model.*50(5),2010, pp. 716-732

[2] S. Agarwal, P. Niyogi, Generalization bounds for ranking algorithms via algorithmic stability, *J. Mach. Learn. Res.,*10,2009, pp. 441-474

[3] M. A. Arcones, A Bernstein-type inequality for U-statistics and U-processes, *Statistics & Probab. Letters,*22, 1995, pp. 239-247

[4] N. Aronszajn, Theory of reproducing kernels, *Trans. Amer. Math. Soc.*68, 1950,pp. 337-404

[5] J. Baldeaux, J. Dick, QMC rules of arbitrary high order: reproducing kernel Hilbert space approach,*Constr. Approx.*30, 2009,pp. 495-527

[6] H. H. Bauschke, P. L. Combettes, *Convex analysis and monotone operator theory in Hilbert spaces,* Springer, New York,2010

[7] A. Caponnetto, E. De Vito, Optimal rates for regularized least-squares algorithm, *Found. Comput. Math.,*7, 2007, pp.331-368

[8] C. Carmeli, E. De Vito, A. Toigo, V.Umanità, Vector valued reproducing kernel Hilbert spaces and university,*Anal. Appl.(Singap),*8(1), 2010,pp. 19-61

[9] C. Carmeli, E. De Vito,A. Toigo,Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem,*Anal. Appl.(Singap),*4(4), 2006,pp. 377-408

[10] J.Chai, J.N.K.Liu, Domainance-based decision rule indunction for multicriteria ranking,*Int. J. Mach. Learn. & Cyber,*4, 2013, pp. 427-444

[11] H. Chen,The convergence rate of regularized ranking algorithm,*J. Approx. Theory,* 164, 2012,pp.1513-1519

[12] H. Chen,Z. B. Pan, L. Q. Li, Learning performance of coefficient-based regularized ranking, *Neurocomputing,* 133, 2014, pp.54-62

[13] H. Chen and J. T. Wu, Regularized ranking with convex losses and l^1-penalty,*Abstr. Appl. Anal.,* Volume 2013, Article ID 927827, 8 pages http://dx. doi. org/10.1155/2013/927827

[14] H. Chen, Y. Tang, L. Q. Li,Y. Yuan, X.L. Li, Y. Y.Tang, Error analysis of stochastic gradient descent ranking, *IEEE Trans. Cybernetics,*43(3), 2013,pp. 898-909

[15] H. Chen, J. T. Peng, Y.C. Zhou, L.Q. Li, Z.B. Pan, Extreme learning machine for ranking: generalization analysis and applications, *Neural Networks,*53, 2014,pp. 119-126

[16] F. Cucker and D. X. Zhou, Learning theory: an approximation theory viewpoint, Cambridge University Press, 2007

[17] Y. Freund,R. E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting,*J. Comput. System Sci.*55,1997,pp.119-139

[18] F. C. He, H.Chen, Generalization performance of bipartite ranking algorithms with convex losses,*J. Math. Anal. Appl.,*404,2013, pp. 528-536

[19] W. Hoeffding, Probability inequalities for sums of bounded random variables,*J. Amer. Statist.Assoc.,*58,1963,pp. 13-30

[20] F. Y. Kuo,G. W. Wasikowski, H. Woźniakowski, Multivariate L_∞ approxiamtion in the worst case setting over reproducing kernel Hilbert spaces,*J. Approx. Theory,* 152,2008,pp.135-160

[21] T. Pahikkala, A. Airola, M. Stock, B. De Baets, W. Waegeman, Efficient regularized least-squares algorithms for conditional ranking on relational data, *Mach. Learn,* 93, 2013,pp. 321-356

[22] T. Pahikkala, H. Suominen, J. Boberg, T. Salakoski,Transductive ranking via pairwise regularized least-squares, In P. Frasconi, K. Kersting, and K. Tsuda, editors, Workshop on Mining and Learning with Graphs (MLG07), 2007

[23] T. Pahikkala, E. Tsivtsivadze, A. Airola, J. Boberg, and T. Salakoski. Learning to rank with pairwise regularized least-squares. In T. Joachims, H. Li, T.-Y. Liu, and C. Zhai, editors, SIGIR 2007 Workshop on Learning to Rank for Information Retrieval, 2007, pp. 27-33,

[24] Z. Pan, X. You, H. Chen, D.Tao,B. Pang, Generalization performance of magnitude-preserving semi-supervised ranking with graph-based regularization,*Inf. Sci.,*221, 2013,pp. 284-296

[25] V. C. Raykar, R. Duraiswami, and B. Krishapuram, A fast algorithm for learning a ranking function from large-scale data sets,*IEEE Trans. Pattern Anal. Mach. Intell.*30(7), 2008,pp. 1158-1170

[26] W. Rejchel, On ranking and generalization bounds, *J. Mach. Learn. Res.,*13, 2012,pp. 1373-1392

[27] B. H. Sheng, The convergence rates of Shannon sampling learning algorithms, *Sci. China Math.,*55(6),2012, pp. 1243-1256

[28] B.H.Sheng,D.H.Xiang,Bound the learning rates with generalized gradients, *WSEAS Trans. Signal Proc.,* 8(1),2012,pp. 1-10

[29] B. H. Sheng, P. X. Ye,The learning rates of regularized regression based on reproducing kernel Banach spaces,*Abstr. Appl. Anal.,*Volume 2013, Article ID 694181, 10 pages http: //dx. doi. org/10. 1155 /2013/694181

[30] B.H.Sheng,W.K.Yu,P.X.Ye,Y.J. Han, Learning rates of regularized regression with p-loss, *Wseas Trans. on Math.,*12(2), 2013, pp. 189-200

[31] I. Steinwart, A. Christmann, Support vector machines, Springer-Verlag, New York Inc., 2008

[32] X. Tian, D.Tao, X.Hua,X.Wu, Active reranking for web search, *IEEE. Trans. Image Process,*,19, 2010,pp. 805-820

[33] M. E. Tipping, Sparse Bayesian learning and the relevance vector machine,*J. Mach. Learn. Res.,*1, 2001, pp. 211-244

[34] E. Tsivtsivadze T. Pahikkala, A. Airola,J. Boberg,and T. Salakoski, A sparse regularized least-squares preference learning algorithm,*Scandinavian Conference on Artificial Intelligence,*2008,pp. 76-83

[35] M. Wang, L.Hao, D.Tao,K.Lu, X.Wu, Mltimodal graph-based reranking for web image search, *IEEE Trans. Image. Process,*21, 2012,pp.4649-4661

[36] F.X. Wang, H.X. Jin, X. Chang, Relevance vector ranking for information retrieval,*J. Conver. Inform. Tech.,* 5(9),2010,pp. 118-125

[37] H. K. Xu, Inequalities in Banach spaces with applications, *Nonlinear Anal.* 16(12),1991,pp. 1127-1138

[38] Z. B. Xu and G. F. Roach, Characteristic inequalities of uniformly convex and uniformly smooth Banach spaces, *J. Math. Anal. Appl.*157(1),1991,pp.189-210

[39] Y. S. Xu, H.Z. Zhang,Q. G.Zhang, Refinement of operator-valued reproducing kernels, *J. Mach. Learn. Res.,*13(1),2012,pp. 91-136

[40] Y. Q. Zhang, F.L.Cao, Analysis of convergence performance of neural networks ranking algorithm, *Neural Networks,*34,2012,pp. 65-71

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