Bound the Learning Rates with Generalized Gradients

SHENG BAOHUAI
Department of Mathematics,
Shaoxing College of Arts and Sciences
Shaoxing, Zhejiang 312000
P.R. China
e-mail: bhsheng@usx.edu.cn

XIANG DAOHONG
Department of Mathematics
Zhejiang Normal University
Jinhua, Zhejiang 321004
P.R.China
e-mail: daohongxiang@gmail.com

Abstract: This paper considers the error bounds for the coefficient regularized regression schemes associated with Lipschitz loss. Our main goal is to study the convergence rates for this algorithm with non-smooth analysis. We give an explicit expression of the solution with generalized gradients of the loss which induces a capacity independent bound for the sample error. A kind of approximation error is provided with possibility theory.

Key Words: Regularization regression, non-smooth analysis, Lipschitz loss,machine learning,learning rates,generalized gradient.

1 Introduction

Recently, in the community of machine learning, considerably attention has been paid for regularized kernel machine learning due to its great success in fields such as signal processing, feature selection and so on (see e.g. [6, 11, 13]). In the present paper, we shall conduct error analysis for the coefficient regularized regression associated with l_2-regularization and Lipschitz loss.

Let X be a compact metric space, $W = \Re$. $\rho(x,w)$ is a fixed but unknown Borel probability distribution on $Z := X \times W$ which describes the relation between variables $x \in X$ and $w \in W$. It can be factorized into the marginal probability $\rho_X(x)$ and the conditional probability $\rho(w|x)$ of w given x.

Let $V(t) : \Re \to \Re_+$ be a convex loss function. Denoted by $\mathcal{E}_{\rho,V}(f) = \int_Z V(|w - f(x)|)\ d\rho$ the generalization error. The minimizerf_V^* defined by

$$f_V^*(x) := arg \min_f \mathcal{E}_{\rho,V}(f)$$

over all measurable functions is the target which we want to learn. In particular, if $V(t) = t^2$ is the least square loss, f_V^* is exactly the regression function $f_\rho(x) = E(w|x) = \int_W w\ d\rho(w|x)$ (see [11]). For the pinball loss and the ε-insensitive loss the existence of f_V^* is studied in [20] and [29] respectively. Therefore, throughout this paper we assume that f_V^* always exists and is uniqueness. A task of learning theory is to find, from the sample $z = (z_i)_{i=1}^m = ((x_i, w_i))_{i=1}^m \in Z^m$ drawn from independent and identically distributed (i.i.d.) random variable (X_i, W_i) each with the unknown probability distribution $\rho(x,w)$ on Z with $1 \le i \le m$, a function $f_z(x) : X \to W$ such that it is a good approximation of f_V^*.

Usually, the function f_z is chosen from the reproducing kernel Hilbert space(RKHS) generated by a Mercer kernel.

Let $K(x,y) : X \times X \to \Re$ be continuous, symmetric and positive semi-definite, i.e., for any finite set of distinct points $\overline{Y} = \{x_1, x_2, \cdots, x_l\} \subset X$, the matrices $(K(x_i,x_j))_{i,j=1}^l$ are positive semi-definite. Such functions are called Mercer kernels.

The reproducing kernel Hilbert space (RKHS)(see [2]) $\mathcal{H}_K$ associating with a Mercer kernel K is defined to be the closure of the linear span of the set of functions $\{K_x := K(x,\cdot) : x \in X\}$ with the inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}_K}$ satisfying

$$\langle K_x, K_y\rangle_{\mathcal{H}_K} = K(x,\ y), \qquad x, y \in X.$$

One way to learn f_z is the following Tikhonov regularization regression (see e.g. [17])

$$f_{z,\lambda,V} = arg \min_{f\in\mathcal{H}_K} \left\{ \frac{1}{m}\sum_{i=1}^m V\Big(|w_i - f(x_i)|\Big) + \lambda \|f\|_{\mathcal{H}_K} \right\}, \quad (1)$$

where λ is a positive constant which is called the regularization parameter. Different penalty functions induce various algorithms. So far many theoretical analysis has been done for scheme (1). The capacity dependent (see e.g. [3, 4, 5, 12, 32, 33]) and the capacity independent (see e.g. [7, 8, 13, 14, 26], et al) are two approaches for this purpose.

Let $V(t) = V_{pb}(t)$ be the pinball loss function defined as (see [31])

$$V_{pb}(t) = \begin{cases} (\tau-1)t, & \text{if } t < 0, \\ \tau t, & \text{if } t \geq 0, \end{cases}$$

where $\tau \in (0,1)$ is a given real number. Then, [7] showed that, if $\lambda = \lambda(m) \to 0$, and $\lambda^2 m \to +\infty$, then,

$$\lim_{m\to+\infty} \|f_{z,\lambda,V_{pb}} - f_V^*\|_0 = 0 \tag{2}$$

holds for all distributions ρ on Z with $|\rho|_1 = \int_Z |y|\, d\rho < +\infty$ and $\|f\|_0 = \int_X \min\{1, |f|\} d\rho_X$.

(2) is a qualitatively description of the convergence of the algorithm (1) when $V(t) = V_{pb}(t)$. It holds for the least square loss and even the p-loss(see [8]). However, the quantitative description for this convergence has not been fully studied. Besides the least square loss the explicit learning rates have not been provided. A reason is that the learning rates are related to RKHS approximation problem which has not been studied fully. In the present paper, we shall give an estimate for the RKHS approximation problem and with which show the learning rates. The fact that $\mathcal{H}_K$ is a infinite dimensional space makes the discussion inconvenient. We shall simplify the framework to a finite dimensional optimization problem and, then, give an explicit learning rate for the learning rates in the case of Lipschitz loss.

By representer theorem we know $f_{z,\lambda,V}(x)$ has the following form (see e.g.[1, 15, 16, 19])

$$f_{z,\lambda,V}(x) = \sum_{j=1}^{m} \alpha_j\, K(x, x_j), \qquad x \in X,$$

for real coefficient vectors $\alpha = (\alpha_1,\ \alpha_2,\ \cdots,\ \alpha_m)^\top$.

Formulation (1) then can be simplified. In fact,[34] defined the following scheme

$$\alpha_{z,\lambda,V} = arg \min_{\alpha\in\Re^m} \{\frac{1}{m}\sum_{i=1}^{m} V(|w_i - f_\alpha(x_i)|) + \lambda\Omega(\alpha)\}, \tag{3}$$

where $f_\alpha \in \mathcal{H}_{K,\overline{X}} = \{f_\alpha(x) = \sum_{j=1}^{m} \alpha_j\ K(x, x_j) : \alpha = (\alpha_1, \alpha_2, \cdots, \alpha_m) \in \Re^m\}$, $\overline{X} = \{x_1, x_2, \cdots, x_m\} \subset X$ is a given data and $\Omega(\alpha) : \Re^m \to [0, +\infty)$ are non-negative functions. In particular, when $\Omega(\alpha) = m\sum_{i=1}^{m} |\alpha_i|^2$, we have

$$\alpha_z^{(V)} : = \arg \min_{\alpha\in\Re^m} \{\frac{1}{m}\sum_{i=1}^{m} V(|w_i - f_\alpha(x_i)|) + \lambda m \sum_{i=1}^{m} \alpha_i^2\}, \tag{4}$$

where $K(x,\ y)$ is a continuous kernel defined on $X \times X$. $\overline{Y} = \{y_1, y_2, \cdots, y_m\} \subset X$ is a discrete subset, $f_\alpha \in \mathcal{H}_{K,\overline{Y}} = \{f_\alpha(x) = \sum_{j=1}^{m} \alpha_j\ K(x, y_j) : x \in X, \alpha = (\alpha_1, \alpha_2, \cdots, \alpha_m) \in \Re^m\}$.

(4) is usually called the coefficient regularized framework (see e.g.[34]) since the regularizer is determined by the coefficient vector α. When $V(t) = t^2$ is the least square loss, the consistency of scheme (4) is studied respectively in [30] and [22] with different methods. In the present paper,we shall provide a kind of learning rates for the mean error

$$\mathcal{E}_{\rho,V}(f_{\alpha_z^{(V)}}) - \mathcal{E}_{\rho,V}(f_V^*)$$

when $V(t)$ is a Lipschitzian and $\overline{Y} \subset X$ is a given discrete subset. We now restate the definition of Lipschitz loss.

Definition 1. *We call a convex function $V(t) : \Re \to \Re_+$ a Lipschitz loss with rank $L > 0$ if it satisfies $V(0) = 0$ and for any $t,\ t' \in \Re$, there holds*

$$|V(t) - V(t')| \leq L|t - t'|.$$

For example, the absolute value loss $V_a(t) = |t|$ is a Lipschitz with rank 1. The $\varepsilon-$insensitive loss

$$V_{il}(t) = \begin{cases} 0, & \text{if } |t| < \varepsilon, \\ |t| - \varepsilon, & \text{otherwise} \end{cases}$$

is a Lipschitz loss with rank 1 (see [21]); By [31] we know pinball loss V_{pb} is a Lipschitz loss with rank τ.

We assume further that

$$k = \sup_{(x,\ y)\in X\times X} |K(x,\ y)| < +\infty. \tag{5}$$

The algorithm (4) considered in this paper has following features:

- The discrete set $\overline{Y}$ may be given ahead according to ours needs. For example, we can choose $\overline{Y}$ such that it is dense in X when $m \to +\infty$ or is taken from the sample, i.e., $\overline{Y} = \overline{X}$. The former is sample independent and the latter is sample dependent. We shall see that whether or not $\overline{Y}$ is sample independent or sample dependent will influence the approximation error.

- Since (4) is a finite dimensional convex optimization problem on $\Re^m$, it is convenient for us to design algorithm.

- In our analysis we only require the loss functions are Lipschitzians, they lack of strong convexity and smoothness, the integral operator approach cannot be used. However, we can use the

Clarke's directional derivative and the generalized gradient to conduct the analysis. These facts make it possible for us to use non-smooth analysis skill in our analysis.

- Since the least squares loss is local Lipschitzian and not the Lipschitzian on $\Re$, the method used here will not suit to the least square loss.

Basing on above facts we shall develop ours approach.

2 Main Results

Before stating our discussions, let us introduce some definitions and notations. The noise free form corresponding to scheme (4) is defined by

$$\alpha^{(\rho)} = \arg\min_{\alpha\in\Re^m} \{\mathcal{E}_{\rho,V}(f_\alpha) + \lambda\, m \sum_{i=1}^{m} \alpha_i^2\}, \qquad (6)$$

where $f_\alpha(x) = \sum_{j=1}^{m} \alpha_j K(x, y_j)$. The corresponding empirical measure $\rho_z(x, w)$ for a bounded $\rho-$measurable function $f(x, w)$ on Z is defined as

$$\int_Z f(x,\ w)\, d\rho_z = \frac{1}{m}\sum_{i=1}^{m} f(x_i, w_i). \qquad (7)$$

We decompose the excess generalization error $\mathcal{E}_{\rho,V}(f_{\alpha_z^{(V)}}) - \mathcal{E}_{\rho,V}(f_V^*)$ into two parts. One is the sample error $|\mathcal{E}_{\rho,V}(f_{\alpha_z^{(V)}}) - \mathcal{E}_{\rho,V}(f_{\alpha^{(\rho)}})|$, the other one is the approximation error $\mathcal{E}_{\rho,V}(f_{\alpha^{(\rho)}}) - \mathcal{E}_{\rho,V}(f_V^*)$. The latter is an approximation problem which will be discussed in Section 3. The former is determined by the loss function and the sample z, whose estimate is the main goal of Section 3.

We present our sample error estimate as following Theorem 1.

Theorem 1. *Let V be a Lipschitz loss with rank L. Assume that $K(x, y)$ satisfies (5). $\alpha^{(\rho)}$ and $\alpha_z^{(V)}$ are defined in (6) and (4) respectively for a given discrete set $\overline{Y} \subset X$. Then, for any $0 < \delta < 1$, with confidence $1 - \delta$, there holds*

$$|\mathcal{E}_{\rho,V}(f_{\alpha_z^{(V)}}) - \mathcal{E}_{\rho,V}(f_{\alpha^{(\rho)}})| \le \frac{4k^2L^2(4log\frac{2}{\delta} + \sqrt{m})\ log\frac{2}{\delta}}{\lambda m}. \qquad (8)$$

(8) shows that the sample error of the regularized regression with Lipschits loss has the same convergence rates as the ones in [26] obtained by the integral operator approach. It leads to the convergence relation:

If $\lambda = \lambda(m) \to 0,$ and $m\lambda^2 \to +\infty$, when $m \to +\infty$,

then, with possibility 1, holds

$$\lim_{m\to+\infty} |\mathcal{E}_{\rho,V}(f_{\alpha_z^{(V)}}) - \mathcal{E}_{\rho,V}(f_{\alpha^{(\rho)}})| = 0.$$

As we shall see in Section 3 that the method used here induces a capacity independent estimate. It is a covering number independent approach and there is no need to make additional assumptions on the kernel space. The method in our arguments has some properties similar to the integral operator approach. For example, both the approaches can gives the explicit expression of the solution. However, the integral operator approach is unsuitable to the Lipschits loss. [7, 8] used the sub-gradient of the loss to describe the convergence for scheme (1). Our arguments will absorb the advantages of [7, 8] and give a kind of learning rates quantitatively.

When the loss function in Theorem 1 become some concrete loss functions, we have the following Corollary 1.

Corollary 1. *If we take the place of loss $V(t)$ in scheme (4) with the pinball loss, the absolute loss or the $\varepsilon-$insensitive loss and assume the kernel $K(x, y)$ satisfies (5), $\alpha_z^{(V)}$ is the uniquely minimizer of scheme (4), then, for any $0 < \delta < 1$, with confidence $1 - \delta$, there holds*

$$|\mathcal{E}_{\rho,V}(f_{\alpha_z^{(V)}}) - \mathcal{E}_{\rho,V}(f_{\alpha^{(\rho)}})| \le \frac{4k^2(4log\frac{2}{\delta} + \sqrt{m})\ log\frac{2}{\delta}}{\lambda m}.$$

We now give an explicit learning rates for learning algorithm (4).

Theorem 2. *Under the conditions of Theorem 1, if $f_V^*(x) = \int_X \varphi(y)\ K(x, y)\ d\rho_X(y)$ and $\varphi \in L^2(\rho_X)$, then, there is a discrete set $\overline{Y} = \{y_1, y_2, \cdots, y_m\} \subset X$ such that the solution $\alpha_z^{(V)}$ of the scheme (4) satisfies, for any $0 < \delta < 1$, with confidence $1 - \delta$, there holds*

$$|\mathcal{E}_{\rho,V}(f_{\alpha_z^{(V)}}) - \mathcal{E}_{\rho,V}(f_V^*)| \le \frac{4k^2\ L^2(4log\frac{2}{\delta} + \sqrt{m})log\frac{2}{\delta}}{\lambda\, m} + L\sqrt{\frac{A - \|f_V^*\|^2_{L^2(\rho_X)}}{m}} + \lambda\, \|\varphi\|^2_{L^2(\rho_X)}, \qquad (9)$$

where $A = \int_X \int_X \ \varphi(y)^2 \ K(x,y)^2 \ d\rho_X(x) \ d\rho_X(y)$ *and* $L^2(\rho_X) = \{f : \|f|_{L^2(\rho_X)} = (\int_X |f(x)|^2 d\rho_X)^{\frac{1}{2}} < +\infty\}$.

Define an integral operator L_K given by

$$L_K(f,x) = \int_X K(x,t) f(t) \ d\rho_X(t).$$

Then, $f_V^*(x) = \int_X \varphi(y) K(x, \ y) \ d\rho_X(y)$ and $\varphi \in L^2(\rho_X)$ implies $f_V^*(x)$ belongs to the range of operator L_K. This type of function class was studied by [26, 27] and in this case we say $f_V^*(x)$ satisfies regularity condition (see [28]).

The rest of the paper is organized as follows. We shall give in Section 3.1 some notations on non-smooth analysis. In Section 3.2, we shall show the robustness for scheme (6), i.e., for any two Borel possibility distributions ρ and γ on Z we give an upper estimate for $\|\alpha^{(\rho)} - \alpha^{(\gamma)}\|$. Theorem 1 is proved in Section 3.3. In Section 4, we shall first prove an estimate for the approximation error, with which and (8) show Theorem 2.

To make the statement clearer, we give here some particular symbols. We shall denote by $K_{\overline{Y}}(x)$ the vector $(K(x,y_1), \cdots, K(x,y_m))$ relating to $\overline{Y}$. By $\Re^m (m \geq 1)$ we denote the $m-$dimensional Euclidean space with the usual inner product, i.e., for any $a = (a_1, a_2, \cdots, a_m)^\top, b = (b_1, b_2, \cdots, b_m)^\top \in \Re^m$, we define

$$\|a\|^2 = \sum_{i=1}^m |a_i|^2 = a^\top \ a, \ \langle a, b\rangle = \sum_{i=1}^m a_i b_i = a^\top b.$$

For a vector function

$$f(x) = (f_1(x), \cdots, f_m(x))^\top$$

and a real function $\alpha(x)$ on X we define

$$f(x)\alpha(x) = (f_1(x)\alpha(x), \cdots, f_m(x)\alpha(x))^\top$$

and

$$\int_X f(x)\alpha(x) d\rho_X = (\int_X f_1(x)\alpha(x) d\rho_X, \cdots, \int_X f_m(x)\alpha(x) d\rho_X)^\top.$$

3 Sample Error Analysis

We know that non-smooth analysis is an important tool in dealing with non-differential optimization problem. We give here some notations on it for the latter needs.

3.1 The generalized gradient

Let f be a Lipschitz function defined on a Hilbert space $(X, \|\cdot\|)$, and let l be any other given vectors in X. The Clarke's generalized directional derivative (see [10]) of f at $x \in X$ in the direction l, denoted by $f^o(x;l)$, is defined as

$$f^o(x; \ l) = \limsup_{x' \to x, \ t \downarrow 0} \frac{f(x'+tl) - f(x')}{t},$$

where of course $x' \in X$ and t is a positive scalar. The generalized gradient of f at x, denoted $\partial f(x)$, is a subset of X defined by

$$\{\xi \in X : f^o(x;l) \geq \langle \xi, l\rangle \qquad \text{for all } l \text{ in } X\},$$

where $\langle \cdot, \cdot \rangle$ denote the inner product produced by the norm $\|\cdot\|$.

If $f(x)$ is a convex function on X, then for any $x' \in X$ there is (see [10]),

$$\partial f(x) = \{\xi \in X : f(x') - f(x) \geq \langle \xi, x' - x\rangle\}, \quad (10)$$

If $f(x) : \Re^m \to \Re$ is a differentiable function, then, $\partial f(x) = \{\nabla f(x)\}$, where $\nabla f(x) = (\frac{\partial f}{\partial x_1}, \ \cdots, \ \frac{\partial f}{\partial x_m})$ is the usual gradient.

A well known result is, if $f(x)$ is a convex function on X, then x_0 is the minimal value point of $f(x)$ on X if and only if $0 \in \partial f(x_0)$(see [10]).

3.2 The robustness

It is not difficult to see that the solutions of the (6) is influenced by the distribution ρ. We call the quantitatively description of this influence the robustness. We shall express the solutions with generalized gradients of the loss and then show the robustness.

Proposition 1. *Let $V(t)$ be a convex loss function on $\Re$, ρ and γ be given Borel probability distributions on Z. $\alpha^{(\rho)}$ and $\alpha^{(\gamma)}$ are the solutions of scheme (6) for ρ and γ respectively. Then, there is an $H(t) \in \partial V(|t|)$ such that*

$$\begin{aligned} & \|\alpha^{(\rho)} - \alpha^{(\gamma)}\| \\ \leq \ & \frac{1}{\lambda m} \| \int_Z H(|w - f_{\alpha^{(\rho)}}(x)|) \ K_{\overline{Y}}(x)^\top \ d\rho \\ & - \int_Z H(|w - f_{\alpha^{(\rho)}}(x)|) \ K_{\overline{Y}}(x)^\top \ d\gamma \|. \end{aligned} \quad (11)$$

(11) shows how the solutions of the scheme (6) is influenced by distribution ρ. In fact, we have the following clearer one.

If $V(t)$ is an even differentiable convex function on $\Re$, then, $\partial V(t) = \{V'(t)\}$, (11) becomes

$$\begin{aligned} & \|\alpha^{(\rho)} - \alpha^{(\gamma)}\| \\ \leq & \frac{1}{\lambda m}\| \int_Z V'(w - f_{\alpha^{(\rho)}}(x))\ K_{\overline{Y}}(x)^\top\ d\rho \\ & - \int_Z\ V'(w - f_{\alpha^{(\rho)}}(x)) K_{\overline{Y}}(x)^\top\ d\gamma\|. \end{aligned}$$

To show Proposition 1 we need some lemmas.

Lemma 1.*Let $V(t)$ be a convex loss function on $\Re, \rho$ be a given Borel probability distribution on Z.$\alpha^{(\rho)}$ is the unique solution of scheme (6) for ρ. Then,*

(i). For any $(x, w) \in Z$ there holds

$$\partial_\alpha V(|w - f_\alpha(x)|) = -\partial V(|w - f_\alpha(x)|) K_{\overline{Y}}(x)^\top. \quad (12)$$

(ii). There is an $H(t) \in \partial V(|t|)$ such that

$$2\lambda m\ \alpha^{(\rho)} = \int_Z\ H(|w - f_{\alpha^{(\rho)}}(x)|)\ K_{\overline{Y}}(x)^\top\ d\rho. \quad (13)$$

Proof. Since $\lambda > 0$, we known $\lambda\|\alpha\|^2$ is a strict convex function about α on $\Re^m$. On the other hand, since $V(t)$ is a convex function, we know (6) is a strict convex optimization problem about α on $\Re^m$. The solutions $\alpha^{(\rho)}$ is thus uniqueness.

Proof of (i). By the Theorem 4.2.1 in [18] we have following result:

If $A : \Re^q \to \Re^q$ is an affine mapping ($Ax = A_0x + b$) with A_0 linear and $b \in \Re^q$ and let g be a finite convex function on $\Re^q$. Then,

$$\partial(g \circ A)(x) = A_0^*\ \partial g(Ax) \quad (14)$$

for all $x \in \Re^q$, where A_0^* is the adjoint of A_0.

Since $V(|w - f_\alpha(x)|) = V(|w - K_{\overline{Y}}(x)\alpha|)$, by taking $A_0 = K_{\overline{Y}}(x)$ and $b = w$ in (14), we have (12).

Proof of (ii). By Proposition 2.2 in [9] we know the following result:

If $f_i\ (i = 1, 2, \cdots, m)$ are Lipschitz on X and let $\lambda_i (i = 1, 2, \cdots, m)$ be scalars. Then, $f(x) = \sum\limits_{i=0}^{m}\ \lambda_i\ f_i(x)$ is Lipschitz on X and we have

$$\partial(\sum_{i=0}^{m} \lambda_i\ f_i(x)) \subset \sum_{i=0}^{m} \lambda_i\ \partial f_i(x). \quad (15)$$

Moreover, if $f_i(x)$ are convex functions on X, then the equality holds.

Since $\int_Z V(|w - f_\alpha(x)|)\ d\rho$ and $\|\alpha\|^2 = \sum\limits_{i=1}^{m} |\alpha_i|^2$ are convex function about α on $\Re^m$, we have by (15) and the fact that $\alpha^{(\rho)}$ is the minimizer of (6) that

$$\begin{aligned} 0 \ \in & \ \partial_\alpha(\int_Z V(|w - f_\alpha(x)|)\ d\rho)|_{\alpha=\alpha^{(\rho)}} + 2\lambda\ m\ \alpha^{(\rho)} \\ = & \int_Z \partial_\alpha V(|w - f_\alpha(x)|)|_{\alpha=\alpha^{(\rho)}}\ d\rho + 2\lambda\ m\ \alpha^{(\rho)} \\ = & -\int_Z\ \partial V(|w - f_{\alpha^{(\rho)}}(x)|)\ K_{\overline{Y}}(x)^\top\ d\rho \\ & + 2\lambda m\ \alpha^{(\rho)}, \end{aligned}$$

where, in the last deduction, we have used the equality (12). Hence, there is $H(t) \in \partial V(|t|)$ such that (13) holds.

Lemma 2. *For any $a, b \in \Re^m$ there holds*

$$\|a\|^2 - \|b\|^2 = 2\langle a - b,\ b\rangle + \|a - b\|^2. \quad (16)$$

(16) can be shown by simple computations.

Proof of (11). Simple computations give

$$f_{\alpha^{(\gamma)}}(x) - f_{\alpha^{(\rho)}}(x) = \langle K_{\overline{Y}}(x)^\top,\ \alpha^{(\gamma)} - \alpha^{(\rho)}\rangle. \quad (17)$$

Let H be defined as in (13). Then, (10) and (17) yields

$$\begin{aligned} & V(|w - f_{\alpha^{(\gamma)}}(x)|) - V(|w - f_{\alpha^{(\rho)}}(x)|) \\ \geq & H(|w - f_{\alpha^{(\rho)}}(x)|)(f_{\alpha^{(\rho)}}(x) - f_{\alpha^{(\gamma)}}(x)) \\ = & \langle \alpha^{(\gamma)} - \alpha^{(\rho)},\ -H(|w - f_{\alpha^{(\rho)}}(x)|) \\ & \times K_{\overline{Y}}(x)^\top\rangle. \end{aligned} \quad (18)$$

(18) is followed by

$$\begin{aligned} & \int_Z\ V(|w - f_{\alpha^{(\gamma)}}(x)|)d\gamma \\ & - \int_Z\ V(|w - f_{\alpha^{(\rho)}}(x)|)\ d\gamma \\ \geq & \langle \alpha^{(\gamma)} - \alpha^{(\rho)},\ -\int_Z\ H(|w - f_{\alpha^{(\rho)}}(x)|) \\ & \times\ K_{\overline{Y}}(x)^\top\ d\gamma\rangle. \end{aligned}$$

Taking $a = \alpha^{(\gamma)}$ and $b = \alpha^{(\rho)}$ into (16), we have

$$\begin{aligned} \|\alpha^{(\gamma)}\|^2 - \|\alpha^{(\rho)}\|^2 \ = & \ 2\langle \alpha^{(\gamma)} - \alpha^{(\rho)}, \alpha^{(\rho)}\rangle \\ & + \|\alpha^{(\gamma)} - \alpha^{(\rho)}\|^2. \end{aligned}$$

Above two equalities yields

$$\begin{aligned} & (\mathcal{E}_{\gamma,V}(f_{\alpha^{(\gamma)}}) + \lambda m\ \|\alpha^{(\gamma)}\|^2) - (\mathcal{E}_{\gamma,V}(f_{\alpha^{(\rho)}}) \\ + \ & \lambda m\ \|\alpha^{(\rho)}\|^2) \end{aligned}$$

$$
\begin{aligned}
&\geq \langle \alpha^{(\gamma)}-\alpha^{(\rho)}, -\int_Z H(|w-f_{\alpha^{(\rho)}}(x)|)K_{\overline{Y}}(x)^\top d\gamma\rangle\\
&\quad +2\lambda m\,\langle \alpha^{(\gamma)}-\alpha^{(\rho)}, \alpha^{(\rho)}\rangle + \lambda m\|\alpha^{(\rho)}-\alpha^{(\gamma)}\|^2\\
&= \langle \alpha^{(\gamma)}-\alpha^{(\rho)}, 2\lambda m\alpha^{(\rho)}\\
&\quad -\int_Z H(|w-f_{\alpha^{(\rho)}}(x)|)K_{\overline{Y}}(x)^\top d\gamma\rangle\\
&\quad +\lambda m\,\|\alpha^{(\rho)}-\alpha^{(\gamma)}\|^2\\
&= \langle \alpha^{(\gamma)}-\alpha^{(\rho)}, \int_Z H(|w-f_{\alpha^{(\rho)}}(x)|)K_{\overline{Y}}(x)^\top d\rho\\
&\quad -\int_Z H(|w-f_{\alpha^{(\rho)}}(x)|)\,K_{\overline{Y}}(x)^\top\,d\gamma\rangle\\
&+ \quad \lambda\, m\,\|\alpha^{(\rho)}-\alpha^{(\gamma)}\|^2, \qquad (19)
\end{aligned}
$$

where, in the last deduction, we have used the equation (13). By the definitions of $\alpha^{(\rho)}$ and $\alpha^{(\gamma)}$ we have

$$
\begin{aligned}
&(\mathcal{E}_{\gamma,V}(f_{\alpha^{(\gamma)}})+\lambda m\|\alpha^{(\gamma)}\|^2)-(\mathcal{E}_{\gamma,V}(f_{\alpha^{(\rho)}})\\
&+\lambda m\,\|\alpha^{(\rho)}\|^2)\leq 0,
\end{aligned}
$$

which and (19) give

$$
\begin{aligned}
&\lambda m\,\|\alpha^{(\rho)}-\alpha^{(\gamma)}\|^2\\
&\leq \langle \alpha^{(\rho)}-\alpha^{(\gamma)}, \int_Z H(|w-f_{\alpha^{(\rho)}}(x)|)\,K_{\overline{Y}}(x)^\top\,d\rho\\
&\quad -\int_Z H(|w-f_{\alpha^{(\rho)}}(x)|)\,K_{\overline{Y}}(x)^\top\,d\gamma\rangle\\
&\leq \|\alpha^{(\rho)}-\alpha^{(\gamma)}\|\times\|\int_Z H(|w-f_{\alpha^{(\rho)}}(x)|)K_{\overline{Y}}(x)^\top\,d\rho\\
&\quad -\int_Z H(|w-f_{\alpha^{(\rho)}}(x)|)\,K_{\overline{Y}}(x)^\top\,d\gamma\|.
\end{aligned}
$$

(11) then holds.

3.3 Proof of Theorem 1

We now show Theorem 1. By equality (16) we have

$$
\begin{aligned}
\|\alpha^{(\rho)}\|^2-\|\alpha_z^{(V)}\|^2 &= 2\langle \alpha^{(\rho)}-\alpha_z^{(V)},\,\alpha_z^{(V)}\rangle\\
&\quad +\|\alpha^{(\rho)}-\alpha_z^{(V)}\|^2. \qquad (20)
\end{aligned}
$$

(20) and the definition of $\alpha^{(\rho)}$ yields

$$
\begin{aligned}
&|\mathcal{E}_{\rho,V}(f_{\alpha^{(\rho)}})-\mathcal{E}_{\rho,V}(f_{\alpha_z^{(V)}})|\\
&\leq \{(\mathcal{E}_{\rho,V}(f_{\alpha_z^{(V)}})+\lambda m\,\|\alpha_z^{(V)}\|^2)\\
&\quad -(\mathcal{E}_{\rho,V}(f_{\alpha^{(\rho)}})+\lambda m\,\|\alpha^{(\rho)}\|^2)\}\\
&\quad +\lambda m\,|\|\alpha^{(\rho)}\|^2-\|\alpha_z^{(V)}\|^2|\\
&= \{(\mathcal{E}_{\rho,V}(f_{\alpha_z^{(V)}})+\lambda m\|\alpha_z^{(V)}\|^2)\\
&\quad -(\mathcal{E}_{\rho,V}(f_{\alpha^{(\rho)}})+\lambda m\,\|\alpha^{(\rho)}\|^2)\}\\
&\quad +\lambda m\,\|\alpha^{(\rho)}-\alpha_z^{(V)}\|\times\|\alpha_z^{(V)}\|\\
&\quad +\lambda m\,\|\alpha^{(\rho)}-\alpha_z^{(V)}\|^2
\end{aligned}
$$

$$
\begin{aligned}
&\leq \{(\mathcal{E}_{\rho,V}(f_{\alpha_z^{(V)}})+\lambda m\,\|\alpha_z^{(V)}\|^2)\\
&\quad -(\mathcal{E}_{\rho,V}(f_{\alpha^{(\rho)}})+\lambda m\,\|\alpha^{(\rho)}\|^2)\}\\
&\quad +2\lambda m\,\|\alpha^{(\rho)}-\alpha_z^{(V)}\|\times\|\alpha_z^{(V)}\|\\
&\quad +\lambda m\,\|\alpha^{(\rho)}-\alpha_z^{(V)}\|^2\\
&= A+2\lambda m\,\|\alpha^{(\rho)}-\alpha_z^{(V)}\|\times\|\alpha_z^{(V)}\|\\
&\quad +\lambda\, m\|\alpha^{(\rho)}-\alpha_z^{(V)}\|^2. \qquad (21)
\end{aligned}
$$

A reformulation of (18) gives

$$
\begin{aligned}
&V(|w-f_{\alpha_z^{(V)}}(x)|)-V(|w-f_{\alpha^{(\rho)}}(x)|)\\
&\leq \langle \alpha_z^{(V)}-\alpha^{(\rho)}, -H(|w-f_{\alpha_z^{(V)}}(x)|)\\
&\quad \times K_{\overline{Y}}(x)^\top\rangle. \qquad (22)
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\mathcal{E}_{\rho,V}(f_{\alpha_z^{(V)}})-\mathcal{E}_{\rho,V}(f_{\alpha^{(\rho)}})\\
&\leq \langle -\int_Z H(|w-f_{\alpha_z^{(V)}}(x)|)\,K_{\overline{Y}}(x)^\top\,d\rho, \alpha_z^{(V)}\\
&\quad -\alpha^{(\rho)}\rangle. \qquad (23)
\end{aligned}
$$

On the other hand, rewrite (20) as

$$
\begin{aligned}
\|\alpha_z^{(V)}\|^2-\|\alpha^{(\rho)}\|^2 &= 2\langle \alpha_z^{(V)}-\alpha^{(\rho)},\,\alpha_z^{(V)}\rangle\\
&\quad -\|\alpha_z^{(V)}-\alpha^{(\rho)}\|^2,
\end{aligned}
$$

which and (22) yields

$$
\begin{aligned}
&A=(\mathcal{E}_{\rho,V}(f_{\alpha_z^{(V)}})+\lambda m\,\|\alpha_z^{(V)}\|^2)-(\mathcal{E}_{\rho,V}(f_{\alpha^{(\rho)}})\\
&+\lambda m\,\|\alpha^{(\rho)}\|^2)\\
&\leq \langle -\int_Z H(|w-f_{\alpha_z^{(V)}}(x)|)\,K_{\overline{Y}}(x)^\top d\rho,\\
&\quad \alpha_z^{(V)}-\alpha^{(\rho)}\rangle+2\lambda m\langle \alpha_z^{(V)},\\
&\quad \alpha_z^{(V)}-\alpha^{(\rho)}\rangle-\lambda m\,\|\alpha_z^{(V)}-\alpha^{(\rho)}\|^2\\
&= \langle -\int_Z H(|w-f_{\alpha_z^{(V)}}(x)|)\,K_{\overline{Y}}(x)^\top\,d\rho\\
&\quad +2\lambda m\,\alpha_z^{(V)},\,\alpha_z^{(V)}-\alpha^{(\rho)}\rangle\\
&\quad -\lambda m\,\|\alpha_z^{(V)}-\alpha^{(\rho)}\|^2.
\end{aligned}
$$

Since (see (13) and (7))

$$
\begin{aligned}
2\lambda m\,\alpha_z^{(V)} &= \frac{1}{m}\sum_{i=1}^m H(|w_i-f_{\alpha_z^{(V)}}(x_i)|)\\
&\quad \times K_{\overline{Y}}(x_i)^\top, \qquad (24)
\end{aligned}
$$

we have

$$
A \leq \langle \frac{1}{m}\sum_{i=1}^m H(|w_i-f_{\alpha_z^{(V)}}(x_i)|)\,K_{\overline{Y}}(x_i)^\top
$$

$$
\begin{aligned}
& -\int_Z H(|w-f_{\alpha_z^{(V)}}(x)|)K_{\overline{Y}}(x)^\top\, d\rho \\
& ,\ \alpha_z^{(V)}-\alpha^{(\rho)}\rangle - \lambda m\, \|\alpha_z^{(V)}-\alpha^{(\rho)}\|^2 \\
\leq\ & \|\int_Z H(|w-f_{\alpha_z^{(V)}}(x)|)\ K_{\overline{Y}}(x)^\top\, d\rho \\
& -\frac{1}{m}\sum_{i=1}^m H(|w_i-f_{\alpha_z^{(V)}}(x_i)|)\ K_{\overline{Y}}(x_i)^\top\| \\
\times\ & \|\alpha_z^{(V)}-\alpha^{(\rho)}\| - \lambda m\, \|\alpha_z^{(V)}-\alpha^{(\rho)}\|^2. \quad (25)
\end{aligned}
$$

(25) and (21) yields

$$
\begin{aligned}
& |\mathcal{E}_{\rho,V}(f_{\alpha^{(\rho)}})-\mathcal{E}_{\rho,V}(f_{\alpha_z^{(V)}})| \\
\leq\ & \|\int_Z H(|w-f_{\alpha_z^{(V)}}(x)|)\ K_{\overline{Y}}(x)^\top\, d\rho \\
& -\frac{1}{m}\sum_{i=1}^m H(|w_i-f_{\alpha_z^{(V)}}(x_i)|)\ K_{\overline{Y}}(x_i)^\top\| \\
& \times\|\alpha_z^{(V)}-\alpha^{(\rho)}\| + 2\lambda\, m\, \|\alpha^{(\rho)}-\alpha_z^{(V)}\| \\
& \times\|\alpha_z^{(V)}\| \\
\leq\ & \Big(\|\int_Z H(|w-f_{\alpha_z^{(V)}}(x)|)\ K_{\overline{Y}}(x)^\top\, d\rho \\
& -\frac{1}{m}\sum_{i=1}^m H(|w_i-f_{\alpha_z^{(V)}}(x_i)|)\ K_{\overline{Y}}(x_i)^\top\| \\
& +2\lambda m\, \|\alpha_z^{(V)}\|\Big)\|\alpha^{(\rho)}-\alpha_z^{(V)}\| \\
=\ & \Big(\|\int_Z H(|w-f_{\alpha_z^{(V)}}(x)|)\ K_{\overline{Y}}(x)^\top\, d\rho \\
& -\frac{1}{m}\sum_{i=1}^m H(|w_i-f_{\alpha_z^{(V)}}(x_i)|)\ K_{\overline{Y}}(x_i)^\top\| \\
+\ & \|\frac{1}{m}\sum_{i=1}^m H(|w_i-f_{\alpha_z^{(V)}}(x_i)|)\ K_{\overline{Y}}(x_i)^\top\|\Big) \\
& \times\|\alpha^{(\rho)}-\alpha_z^{(V)}\|,
\end{aligned}
$$

where, in the last deduction, we have used the equation (24).

On the other hand, (7) and (11) yields

$$
\begin{aligned}
& \|\alpha^{(\rho)}-\alpha_z^{(V)}\| \\
& \leq \frac{1}{\lambda m}\|\int_Z H(|w-f_{\alpha_z^{(V)}}(x)|)\ K_{\overline{Y}}(x)^\top\, d\rho \\
& -\frac{1}{m}\sum_{i=1}^m H(|w_i-f_{\alpha_z^{(V)}}(x_i)|)\ K_{\overline{Y}}(x_i)^\top\|.
\end{aligned}
$$

We then have

$$
\begin{aligned}
& |\mathcal{E}_{\rho,V}(f_{\alpha^{(\rho)}})-\mathcal{E}_{\rho,V}(f_{\alpha_z^{(V)}})| \\
\leq\ & \frac{1}{\lambda m}\Big(\|\int_Z H(|w-f_{\alpha_z^{(V)}}(x)|)\ K_{\overline{Y}}(x)^\top\, d\rho \\
& -\frac{1}{m}\sum_{i=1}^m H(|w_i-f_{\alpha_z^{(V)}}(x_i)|)\ K_{\overline{Y}}(x_i)^\top\| \\
+\ & \|\frac{1}{m}\sum_{i=1}^m H(|w_i-f_{\alpha_z^{(V)}}(x_i)|)K_{\overline{Y}}(x_i)^\top\|\Big) \\
& \times\|\int_Z H(|w-f_{\alpha_z^{(V)}}(x)|)\ K_{\overline{Y}}(x)^\top\, d\rho \\
& -\frac{1}{m}\sum_{i=1}^m H(|w_i-f_{\alpha_z^{(V)}}(x_i)|) \\
& \times K_{\overline{Y}}(x_i)^\top\|. \quad (26)
\end{aligned}
$$

By the Proposition 1.5 in Chapter 2 of [9] we know

If $f(x):\Re\to\Re$ is a Lipschitz function with rank L, then, $\partial f(x)$ is a nonempty, convex, subset of $\Re$, and $\|\xi\|\leq L$ for every $\xi\in\partial f(x)$.

Then, the assumption that $V(t)$ is a Lipschitz loss with rank L yields $|H(|w-f_{\alpha_z^{(V)}}(x)|)|\leq L$ for all $(x,\ w)\in X\times W$, which and the fact $\|K_{\overline{Y}}(x)^\top\| = (\sum_{j=1}^m |K(x,\ y_j)|^2)^{\frac{1}{2}} \leq k\sqrt{m}$ ensures

$$
\begin{aligned}
& \|H(|w-f_{\alpha_z^{(V)}}(x)|)\ K_{\overline{Y}}(x)^\top\| \\
=\ & |H(|w-f_{\alpha_z^{(V)}}(x)|)|\times\|K_{\overline{Y}}(x)^\top\| \\
\leq\ & kL\sqrt{m} \quad (27)
\end{aligned}
$$

and

$$
\begin{aligned}
& \int_Z \|H(|w-f_{\alpha_z^{(V)}}(x)|)\ K_{\overline{Y}}(x)^\top\ \|^2\, d\rho \\
& \leq k^2L^2m. \quad (28)
\end{aligned}
$$

Notice the following large number law(see [26]):

Let $(H,\|\cdot\|_H)$ be a Hilbert space and ξ be a random variable on (Z,ρ) with values in H. Assume $\|\xi\|_H\leq\widetilde{M}<+\infty$ almost surely. Denoted $\sigma^2(\xi)=E(\|\xi\|_H^2)$. Let $\{\xi_i\}_{i=1}^m$ be independent random drawers of ρ. For any $0<\delta<1$, with confidence $1-\delta$, there holds

$$
\begin{aligned}
& \|\frac{1}{m}\sum_{i=1}^m(\xi_i-E(\xi_i))\|_H \\
& \leq \frac{2\widetilde{M}log(2/\delta)}{m}+\sqrt{\frac{2\sigma^2(\xi)\ log(2/\delta)}{m}}. \quad (29)
\end{aligned}
$$

Take $\xi(x,w)=H(|w-f_{\alpha_z^{(V)}}(x)|)\ K_{\overline{Y}}(x)^\top$, then, (27) and (28) yields

$$
\|\xi\|\leq kL\sqrt{m} \text{ and } E(\|\xi\|^2)\leq k^2L^2m. \quad (30)
$$

We know by (29) and (30) that, with confidence $1-\delta$, there holds

$$
\|\int_Z H(|w-f_{\alpha_z^{(V)}}(x)|)\ K_{\overline{Y}}(x)^\top\, d\rho
$$

$$
\begin{aligned}
& -\frac{1}{m}\sum_{i=1}^{m} H(|w_i - f_{\alpha_z^{(V)}}(x_i)|)\, K_{\overline{Y}}(x_i)^\top \| \\
\leq & \sup_{\|\xi\|\leq kL\sqrt{m}} \| \int_Z \xi(x,\, w)\, d\rho - \frac{1}{m}\sum_{i=1}^{m} \xi(x_i,\, w_i)\| \\
\leq & kL\sqrt{m} \times (\frac{2log\frac{2}{\delta}}{m} + \sqrt{\frac{2log\frac{2}{\delta}}{m}}) \\
\leq & 4kLlog\frac{2}{\delta}. \qquad (31)
\end{aligned}
$$

Take (31) and (27) into (26) give (8).

4 Approximation Error and Learning Rates

By Section 2 we know, besides the sample error,we need the approximation error to show the learning rates. The approximation error is related to RKHS approximation. When $V(t)$ is the least square loss, the approximation problem has been studied by [27]. For some chosen discrete sets $\overline{Y}$ the approximation problem is investigated in [23, 24, 25]. In this section, we shall give an upper bound for the approximation error with possibility theory.

Proposition 2. *Under the conditions of Theorem 1 if* $f_V^*(x) = \int_X \varphi(y)K(x,\ y)\ d\rho_X(y)$ *and* $\varphi \in L^2(\rho_X) = \{f(x) : \|f\|_{L^2(\rho_X)} = (\int_X |f(x)|^2\ d\rho_X)^{\frac{1}{2}} < +\infty\}$, *then, there is a discrete set* $\overline{Y} \subset X$ *such that the solution* $\alpha^{(\rho)}$ *of scheme (6) satisfies*

$$
\begin{aligned}
& \mathcal{E}_{\rho,V}(f_{\alpha^{(\rho)}}) - \mathcal{E}_{\rho,V}(f_V^*) \\
& \leq L\sqrt{\frac{A - \|f_V^*\|^2_{L^2(\rho_X)}}{m}} + \lambda\|\varphi\|^2_{L^2(\rho_X)}, \qquad (32)
\end{aligned}
$$

where $A = \int_X \int_X \varphi(y)^2\ K(x,\ y)^2 d\rho_X(x)\ d\rho_X(y)$.

Proof. The definition of $\alpha^{(\rho)}$ yields

$$
\begin{aligned}
& \mathcal{E}_{\rho,V}(f_{\alpha^{(\rho)}}) - \mathcal{E}_{\rho,V}(f_V^*) \\
\leq & \mathcal{E}_{\rho,V}(f_{\alpha^{(\rho)}}) - \mathcal{E}_{\rho,V}(f_V^*) + \lambda m\|\alpha^{(\rho)}\|^2 \\
= & \inf_{\alpha\in\Re^m} (\mathcal{E}_{\rho,V}(f_\alpha) - \mathcal{E}_{\rho,V}(f_V^*) + \lambda m\|\alpha\|^2).
\end{aligned}
$$

Since $V(t)$ is a Lipschitz loss with rank L, we have

$$
\begin{aligned}
& \mathcal{E}_{\rho,V}(f_\alpha) - \mathcal{E}_{\rho,V}(f_V^*) \\
= & |\mathcal{E}_{\rho,V}(f_\alpha) - \mathcal{E}_{\rho,V}(f_V^*)| \\
\leq & \int_Z |V(|f_\alpha(x) - w|) - V(|f_V^*(x) - w|)|\ d\rho \\
\leq & L\int_X |f_\alpha(x) - f_V^*(x)|\ d\rho_X.
\end{aligned}
$$

Let $\{X_1, X_2, \cdots, X_m\}$ be i.i.d. random variables with the same distribution ρ_X and for a function $g(X_1,\ X_2,\ \cdots,\ X_m)$ on X^m we define the mathematical expectation $E(g)$ as

$$
\begin{aligned}
E(g) = & \int_{X^m} g(X_1,\ X_2,\ \cdots,\ X_m)\ d\rho_X(X_1) \\
& \times \cdots\ d\rho_X(X_m),
\end{aligned}
$$

then, there is $\overline{Y} = (y_1, y_2, \cdots, y_m) \subset X$ such that

$$
\begin{aligned}
& \mathcal{E}_{\rho,V}(f_{\alpha^{(\rho)}}) - \mathcal{E}_{\rho,V}(f_V^*) \\
\leq & \mathcal{E}_{\rho,V}(f_{\alpha^{(\rho)}}) - \mathcal{E}_{\rho,V}(f_V^*) \\
& +\lambda m\ \|\alpha^{(\rho)}\|^2 \\
\leq & \inf_{\alpha\in\Re^m} (L\int_X |f_\alpha(x) - f_V^*(x)|\ d\rho_X + \lambda m\ \|\alpha\|^2) \\
\leq & L\int_X \Big|\sum_{k=1}^{m} \frac{\varphi(y_k)}{m} \times K(x, y_k) - f_V^*(x)\Big|\ d\rho_X \\
& +\lambda \sum_{k=1}^{m} \frac{\varphi(y_k)^2}{m} \\
\leq & E(L\int_X \Big|\sum_{k=1}^{m} \frac{\varphi(X_k)}{m} \times K(x, X_k) - f_V^*(x)\Big|\ d\rho_X \\
& +\lambda \sum_{k=1}^{m} \frac{\varphi(X_k)^2}{m}) \\
\leq & \lambda\|\varphi\|^2_{L^2(\rho_X)} \\
& +L\int_X E(|\sum_{k=1}^{m} \frac{\varphi(X_k)}{m} \times K(x,\ X_k) - f_V^*(x)|)\ d\rho_X.
\end{aligned}
$$

The Cauchy's inequality gives

$$
\begin{aligned}
& E[|\sum_{k=1}^{m} \frac{\varphi(X_k)}{m} \times K(x,\ X_k) - f_V^*(x)|] \\
& \leq (E(|\sum_{k=1}^{m} \frac{\varphi(X_k)}{m} \times K(x, X_k) - f_V^*(x)|^2))^{\frac{1}{2}}.
\end{aligned}
$$

It follows by the Hölder inequality that

$$
\begin{aligned}
& \int_X E(|\sum_{k=1}^{m} \frac{\varphi(X_k)}{m} \times K(x,\ X_k) - f_V^*(x)|) \\
& \times d\rho_X(x) \\
\leq & \int_X (E[|\sum_{k=1}^{m} \frac{\varphi(X_k)}{m} \times K(x,\ X_k) - f_V^*(x)|^2])^{\frac{1}{2}} \\
& \times d\rho_X(x) \\
\leq & (\int_X E[|\sum_{k=1}^{m} \frac{\varphi(X_k)}{m} \times K(x,\ X_k) - f_V^*(x)|^2] \\
& \times d\rho_X(x))^{\frac{1}{2}}.
\end{aligned}
$$

Therefore,

$$\begin{aligned}
&\mathcal{E}_{\rho,V}(f_{\alpha^{(\rho)}}) - \mathcal{E}_{\rho,V}(f_V^*)\\
\leq\ & \lambda \|\varphi\|^2_{L^2(\rho_X)}\\
+\ & L\Big(\int_X E(|\sum_{k=1}^m \frac{\varphi(X_k)}{m} \times K(x,X_k) - f_V^*(x)|^2)\\
& \times d\rho_X\Big)^{\frac{1}{2}}\\
=\ & \lambda\|\varphi\|^2_{L^2(\rho_X)} + L\Big(\int_X E(|f_V^*(x)|^2\\
& -2\sum_{k=1}^m \frac{\varphi(X_k)\ K(x,\ X_k)}{m} \times f_V^*(x)\\
& + \sum_{k,j=1}^m \frac{\varphi(X_k)\ \varphi(X_j)}{m^2} \times K(x,\ X_k)\\
& \times K(x,\ X_j))d\rho_X\Big)^{\frac{1}{2}}. \qquad (33)
\end{aligned}$$

Since $X_1, X_2, \cdots, X_m$ are independent and

$$f_V^*(x) = \int_X \varphi(y)K(x,y)\ d\rho_X(y),$$

we have

$$\begin{aligned}
& E\Big(|f_V^*(x)|^2 - 2\sum_{k=1}^m \frac{\varphi(X_k)\ K(x,X_k)}{m} \times f_V^*(x)\\
& + \sum_{k,j=1}^m \frac{\varphi(X_k)\varphi(X_j)}{m^2} \times K(x,\ X_k)\ K(x,X_j)\Big)\\
=\ & |f_V^*(x)|^2 - \frac{2}{m}\sum_{k=1}^m E(\varphi(X_k)K(x,X_k)) \times f_V^*(x)\\
& + \frac{1}{m^2}\sum_{k,j=1}^m E(\varphi(X_k)\varphi(X_j)\ K(x,\ X_k)\ K(x,\ X_j))\\
=\ & -|f_V^*(x)|^2\\
& + \frac{1}{m^2}\sum_{k=j} E\Big[\varphi(X_k)\varphi(X_j)K(x,X_k)\\
& \times K(x,X_j)\Big]\\
& + \frac{1}{m^2}\sum_{k\neq j} E\Big[\varphi(X_k)\ \varphi(X_j)K(x,X_k)K(x,X_j)\Big]\\
=\ & -|f_V^*(x)|^2 + \frac{1}{m}\int_X \varphi(y)^2K(x,y)^2\ d\rho_X(y)\\
& + \frac{m(m-1)}{m^2} \times |f_V^*(x)|^2
\end{aligned}$$

and therefore

$$\begin{aligned}
& \int_X E\Big(|f_V^*(x)|^2 - 2\sum_{k=1}^m \frac{\varphi(X_k)\ K(x,\ X_k)}{m}\\
& \times f_V^*(x) + \sum_{k,j=1}^m \frac{\varphi(X_k)\varphi(X_j)}{m^2}\\
& \times K(x,X_k)\ K(x,X_j)\Big)d\rho_X\\
=\ & -\|f_V^*\|^2_{L^2(\rho_X)}\\
& + \frac{1}{m}\int_X\int_X \varphi(y)^2K(x,y)^2\ d\rho_X(y)d\rho_X(x)\\
& + \frac{m(m-1)}{m^2} \times \|f_V^*(x)\|^2_{L^2(\rho_X)}\\
=\ & \frac{\int_{X\times X} \varphi(y)^2K(x,y)^2 d\rho_X(y)d\rho_X(x) - \|f_V^*\|^2_{L^2(\rho_X)}}{m}.
\end{aligned}$$

(33) and above equality yields (32).

Proof of Theorem 2. (9) can be obtained by (32) and (8).

Acknowledgements: The research is supported by NSF of China(under grant numbers 10871226, 11001247) and the NSF of Zhejiang province (under grant numbers Y6100096).

References:

[1] A. Argyriou, C. A. Micchelli,M. Pontil, When is there a representer theorem? vector versus matrix regularizers, *J. Mach. Learn. Res.* 10,2009, pp. 2507-2529

[2] N.Aronszajn, Theory of reproducing kernels. *Trans. Amer. Math. Soc.,* 68(2),1950,pp. 337-404

[3] P. L.Bartlett, S. Mendelson, Rademacher and Gaussian complexities: Risk bounds and structural results, *J. Mach. Learn. Res.,* 3,2002, pp. 463-482

[4] H. Chen,L. Q. Li, J. T. Peng, Semi-supervised learning based on high density region estimation, *Neural Networks,* 23(7),2010, pp.12-818

[5] H. Chen, L. Q. Li, On the rate of convergence for multi-category classification based on convex losses, *Science in China, Series A: Mathematics,* 50 (11), 2007, pp. 1529-1536

[6] A. Christmann, On a strategy to develop robust and simple tariffs from motor vehicle insurance data, *Acta Mathematicae Applicatae Sinica, English Series,* 21(2),2005, pp. 193-208

[7] A. Christmann, I. Steinwart, Consistency of kernel based quantile regression, *Appl. Stoch. Model. Bus. and Industr.,* 24(2),2008, pp.171-183

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

[9] F. H.Clarke, Yu. S. Ledyaev, R. J. Stern, P. R. Wolenski, *Nonsmooth analysis and control theory,* Springer-Verlag, Berlin, 1998

[10] F. H. Clarke, *Optimization and nonsmooth analysis,* John Wiley and Sons, Inc., 1983

[11] F. Cucker, S. Smale, On the mathematical foundations of learning theory, *Bull. Amer. Math.*Soc., 39(1),2001, pp.1-49

[12] F. Cucker,D. X.Zhou, *Learning theory: An approximation theory viewpoint,* Cambridge University Press, New York, 2007

[13] C. De Mol, E. De Vito, L. Rosasco, Elastic-net regularization in learning theory, *J. Complexity,* 25(2),2009,pp. 201-230

[14] E. De Vito, L. Rosasco, L. Caponnetto, M. Piana, A. Verri, Some properties of regularized kernel methods, *J. Mach. Learn. Res.,* 5,2004, pp. 1363-1390

[15] F. Dinuzzo, G. Nicolao, An algebraic characterization of the optimum of regularized kernel methods, *Machine Learning* 74(3),2009, pp. 315 - 345

[16] F.Dinuzzo, M.Neve, G. D. Nicolao,U. P. Gianazza, On the representer theorem and equivalent degrees of freedom of SVR, *J. Mach. Learn. Res.,* 8, 2007, pp. 2467-2495

[17] T. Evgeniou, M. Pontil, T. Poggio, Regularization networks and support vector machines, *Adv. Comput. Math.,* 13(1),2000,pp. 1-50

[18] J.-B.Hiriart-Urruty, C. Lemaréchas, *Fundamental of convex analysis,* Springer-Verlag, Berlin,2001

[19] C. A. Micchelli, M. Pontil,Learning the kernel function via regularization, *J. Mach. Learn. Res.,* 6, 2005, pp.1099-1125

[20] N. Quadrianto, K. Kersting, M. Reid, T. Caetano, et al, Kernel conditional quantile estimation via reduction revisited, *ICDM, Proceedings of the 2009 Ninth IEEE International Conference on Data Mining,*2009 938-943

[21] L. Rosasco, E. De Vito, A. Caponnetto, M. Piana, A. Verri, Are loss function all the same?, *Neural Comput.,* 16(5),2004,pp. 1063-1076

[22] B. H. Sheng, P. X. Ye, Least square regression learning with data dependent hypothesis and coefficient regularization, *J. Computer* 6(4),2011,pp. 671-675

[23] B. H. Sheng, D. H. Xiang, The convergence rate for a K-functional in learning theory, *J. Inequali. Appli., 2010.* Article ID 249507, 18 pages doi: 10. 1155/2010/249507

[24] B. H. Sheng, On approximation by reproducing kernel spaces in weighted L^p spaces., *J. Syst. Sci. and Complexity,* 20(4),2007,pp. 623-638

[25] B. H. Sheng, On the degree of approximation by spherical translation, *Acta Math. Appli. Sinica, English Series,*22(4), 2006, pp. 671-680

[26] S. Smale,D. X.Zhou, Learning theory estimates via integral operators and their applications, *Constr. Approx.,* 26(2), 2007, pp. 153-172

[27] S.Smale, D. X. Zhou, Estimating the approximation error in learning theory, *Anal. and Appl.*1(1), 2003, pp. 17-41

[28] S.Smale, D. X. Zhou, Online learning with Markov sampling,*Anal. and Appli.*7(1),2009, pp.87-113

[29] I. Steinwart, A. Christmann, How SVMs can estimate quantiles and media. *Advances in Neural Information Processing System,* 20(2), 2008, pp.305-312

[30] H. W. Sun, Q. Wu, Least square regression with indefinite kernels and coefficient regularizationn, *Appl. Comput. Harm. Anal.* 30(1),2011, pp. 96-109

[31] I.Takeuchi,Q. V. Le, T. D.Sears, T. D.Smola, Nonparametric quantile estimation, *J. Mach. Learn. Res.,* 7, 2006, pp. 1231-1264

[32] H. Z. Tong, D. R. Chen, L. Z. Peng, Analyis of support vector machine regression. *Found. Comput. Math.,*9(2), 2009, pp.243-257

[33] V. Vapnik, *Statistical learning theory,* John Wiley and Sons, New York, 1998

[34] Q. Wu, D.X. Zhou, Learning with sample dependent hypothesis spaces, *Comput. Math. with Appli.,* 56(11), 2008, pp. 2896-2907.