The Consistency Analysis of Coefficient Regularized Classification with Convex Loss

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: It is known that the learning rates are the quantitative description of the consistency of a learning algorithm. In the present paper, we provide the learning rates for the coefficient regularized classification learning algorithm with a $K-$functional whose explicit rates are estimated when the loss functions are least square loss and the hinge loss.

Key–Words: Coefficient regularized classification, machine learning, convex loss, differentiable loss, K-functional, learning rates.

1 Introduction

In the present paper, we give an investigate on the learning rates of the coefficient regularized classification algorithm with differentiable convex loss.

Let (X, d) be a compact metric space, $W = \{-1, 1\}, \rho(x, w) = \rho(w|x)\rho_X(x)$ be a unknown probability distribution on $Z := X \times W$, where $\rho(w|x)$ is the conditional probability distribution of w for a given x and $\rho_X(x)$ is the marginal probability distribution of x. It is known that a binary classifier is a function $f(x) : X \to W$ which divides X into two classes, its prediction ability is measured by the misclassification error (see [6, 34, 36, 37])

$$\begin{aligned}\Re(f) &= Prob\{f(x) \neq w\} \\ &= \int_X P(w \neq f(x)|x)d\rho_X(x).\end{aligned}$$

By [13] we know the classifier which minimizes the misclassification error is the Bayes rule $f_c := sgn(f_\rho)$ with f_ρ being the regression function of ρ,i.e.,

$$\begin{aligned}f_\rho(x) &= \int_W wd\rho(w|x) \\ &= P(w = 1|x) - P(w = -1|x),\end{aligned}$$

where for a function $f : X \to R$ the sign function is defined as $sgn(f)(x) = 1$ if $f(x) \geq 0$ and $sgn(f)(x) = -1$ if $f(x) < 0$. However, in many practical applications, neither ρ nor f_ρ are known, what we have in hand are samples $z = \{(w_i, w_i)\}_{i=1}^m$ drawn independently according to ρ. The task of classification learning is to find, through the samples z, a good approximation f_z of the regression function f_ρ from a hypothesis space and show the excess misclassification error (see e.g. [6, 11, 30, 37])

$$\Re(sgn(f_z)) - \Re(f_c). \tag{1}$$

The hypothesis space are usually taken to be the reproducing kernel Hilbert spaces.

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

The reproducing kernel Hilbert space(RKHS)(see e.g.[10, 11]) $\mathcal{H}_K$ associated with a Mercer kernel $K(x, y)$ is defined to be the closure of the linear span of the set of functions $\{K_x := K(x, \cdot) : x \in X\}$ with an inner product $\langle\cdot, \cdot\rangle_K$ satisfying

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

The reproducing property takes the form

$$f(x) = \langle f, K_x\rangle_K, \qquad x \in X.$$

A method of finding the binary classifier $f_c(x)$ through the samples z is the SVM Tikhonov regularization classification algorithm.

Let $V(t) : R \to [0, +\infty)$ be a given normalized loss function for classification which will be defined afterwards. Then, the Tikhonov regularized support

vector machine classification learning algorithms associated with the reproducing kernel Hilbert spaces (RKHS) $(\mathcal{H}_K, \|\cdot\|_K)$ are defined by (see e.g. [11, 15])

$$f_z := arg \min_{f\in\mathcal{H}_K} (\frac{1}{m}\sum_{i=1}^{m} V(w_i f(x_i)) + \lambda\|f\|_K), \quad (2)$$

where $1 > \lambda > 0$ are given regularization parameters which are commonly used to overcome the ill-posedness.

By [2] we know the unique solutions $f_z(x)$ of (2) takes the form

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

where $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_m)^\top \in R^m$. Scheme (2) then can be simplified. In fact, [35] defined the following general coefficient regularized scheme

$$\begin{aligned} f_z &= f_{\alpha_z}, \\ \alpha_z &= arg \min_{\alpha\in R^m} [\frac{1}{m}\sum_{i=1}^{m} V(w_i f_\alpha(x_i)) \\ &\quad +\lambda\Omega(\alpha)], \end{aligned} \quad (3)$$

where $\Omega(\alpha) : R^m \to R$ is a non-negative function satisfying $\Omega(0) = 0$, and

$$\begin{aligned} \mathcal{H}_{K,\overline{X}} &= \{f_\alpha(x) = \sum_{j=1}^{m} \alpha_j\ K(x, x_j) : \\ &\quad \alpha = (\alpha_1, \alpha_2, \cdots, \alpha_m)^\top \in R^m\} \end{aligned}$$

are kernel function spaces produced by the samples $\overline{X} = \{x_1, x_2, \cdots, x_m\}$ and the kernels $K(x, y)$. The studies given in [28, 35, 38] show that the error analysis for (3) is not easy since the kernel spaces are dependent upon the samples. On the other hand, we notice that in many cases, for example,in the function reconstruction and variable selection(see [14, 16, 17, 19, 22, 29, 31]), one often takes the data dependent kernel space

$$\begin{aligned} \mathcal{H}_{K,\overline{Y}} = \{f_\alpha(x) &= \sum_{j=1}^{m} \alpha_j\ K(x, y_j) : \\ \alpha &= (\alpha_1, \alpha_2, \cdots, \alpha_m)^\top \in R^m\} \end{aligned}$$

as the hypothesis space, where $\overline{Y} = \{y_1, y_2, \cdots, y_m\}$ is a given data in X. It can be equipped with some properties according to our needs. For example, we can choose $\overline{Y} \subset X$ such that $\mathcal{H}_{K,\overline{Y}}$ is density in the measurable functions space which makes it possible for us to construct the kernel approximating operators (see [23]). When $\Omega(\alpha) = m\sum_{i=1}^{m} |\alpha_i|^2$, we have the following coefficient regularized scheme with l_2-penalization (see also [28, 35])

$$\begin{aligned} \alpha_z : &= \arg \min_{\alpha\in R^m} [\frac{1}{m}\sum_{i=1}^{m} V(w_i f_\alpha(x_i)) \\ &\quad +\lambda m\sum_{i=1}^{m} \alpha_i^2], \end{aligned} \quad (4)$$

where $f_\alpha \in \mathcal{H}_{K,\overline{Y}}$.

Equation (4) is a finite dimensional optimization problem on R^m and its optimal solutions may be obtained by optimal computation algorithm.

To show the performance of the algorithm (4), we need to estimate the error (1). For these purposes we need to study the mean error between $f_{\alpha_z}(x)$ and the ideal estimator $f^*(x)$ defined by (see 10)

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

where $\mathcal{E}_{\rho,V}(f) = \int_Z V(wf(x))d\rho$. The minimum is taken over all the functions which are measurable with respect to ρ_X. If $V(t) = t^2$ is the least square loss, then,f^* is exactly the regression function

$$f_\rho(x) = E(w|x) = \int_Y w d\rho(w|x)$$

and if $V(t)$ is the hinge loss

$$V_h(t) = (1-t)_+ = max\{1-t, 0\},$$

then, $f^* = f_c$ (see [37]).

(4) may be interpreted as a stochastic approximation of the following regularized risk minimization

$$\alpha^{(\rho)} := \alpha_{\lambda,V}^{(\rho)} = arg \min_{\alpha\in R^m} [\mathcal{E}_{\rho,V}(f_\alpha) + \lambda m\|\alpha\|_2^2]. \quad (5)$$

We now define the empirical distribution $\mu_z(x, w)$ on Z by

$$\begin{aligned} E_{\mu_z}[f(x, w)] &= \int_Z f(x, w)d\mu_z \\ &= \frac{1}{m}\sum_{i=1}^{m} f(x_i, w_i) \end{aligned} \quad (6)$$

for any bounded $\rho-$measurable function $f(x, w)$ on Z.

The normalized loss functions for classification are defined as follows:

Definition 1 *(see [30]). A function $V : R \to [0, +\infty)$ is called a normalized loss function for classification if it is convex, $V'(0) < 0, 1$ is the minimal zero of $V(t)$.*

Examples of such normalized loss functions include the hinge loss $V_h(t) = (1-t)_+$ for classical SVM classifier and the $q-$norm $V_q(t) = (1-t)_+^q$ for SVM $q-$norm$(q > 1)$ soft margin classifier(see, e.g. [6]), the least square loss $V_{ls}(t) = (1-t)^2$(see [39]) and other loss functions (see [20]).

Basing on above notations we give the main result of the present paper.

Theorem 1.1. *Let $\rho(x,w)$ be a (joint) finite distribution on $Z, z = \{(x_i,w_i)\}_{i=1}^m$ be samples drawn randomly according to $\rho(x,w)$ independently. $V(t)$ is a given normalized loss functions on R and satisfies $V''(0) \geq 0$.$K(x,y)$ is a given kernel on $X \times X$ and satisfies $k = \sup\limits_{(x,y)\in X\times X} |K(x,y)| < +\infty$. $\overline{Y} = \{y_1, y_2, \cdots, y_m\}$ is a given discrete set in X. α_z is the uniquely minimizer of scheme (4). Then, there is a constant C_V depending upon $V(t)$ such that, for any $0 < \delta < 1$, with confidence $1-\delta$, there holds*

$$\begin{aligned} & 0 \leq \Re(sgn(f_z)) - \Re(f_c) \\ \leq\ & C_V\Big[\frac{4k^2(4log\frac{2}{\delta} + \sqrt{m})\ log\frac{2}{\delta}}{\lambda m} \\ \times\ & \|V'\|^2_{[-k\sqrt{\frac{V(0)}{\lambda}}, k\sqrt{\frac{V(0)}{\lambda}}]} + K_V^{(\overline{Y})}(f^*,\lambda)\Big]^{\frac{1}{2}}, \end{aligned} \tag{7}$$

where $\|V\|_{[a,b]} = ess\sup\limits_{t\in[a,b]} |V(t)|$ and the $K-$functional

$$\begin{aligned} K_V^{(\overline{Y})}(f^*,\lambda) &= \inf_{\alpha\in R^m} (\mathcal{E}_{\rho,V}(f_\alpha) - \mathcal{E}_{\rho,V}(f^*) \\ &\quad + m\lambda\ \|\alpha\|_2^2). \end{aligned}$$

(7) shows that, to give the learning rates,we need to show the explicit convergence rates for the $K-$functional $K_V^{(\overline{Y})}(f^*,\lambda)$. When $V_{ls}(t) = (1-t)^2$ is the least square loss, we have the following estimate.

Theorem 1.2. *Let ρ be a (joint) finite nonnegative distribution on $Z, z = \{(x_i,w_i)\}_{i=1}^m$ be samples drawn randomly and independently according to $\rho(x,w)$. $K(x,y)$ is a given kernel on $X \times X$ and satisfies $k = \sup\limits_{(x,y)\in X\times X} |K(x,y)| < +\infty$. $\overline{Y} = \{y_1, y_2, \cdots, y_m\}$ is a given discrete set in X. α_z is the uniquely minimizer of scheme (4) for the least square loss $V_{ls}(t) = (1-t)^2$. Then, there is a constant $C > 0$ such that, for any $0 < \delta < 1$, with confidence $1-\delta$, there holds*

$$\begin{aligned} & \Re(sgn(f_z)) - \Re(f_c) \\ \leq\ & C[\frac{6k^2\log\frac{2}{\delta}}{\lambda\sqrt{m}} + \sqrt{K_{V_{ls}}^{(\overline{Y}),*}(f_\rho,\lambda)}], \end{aligned} \tag{8}$$

where

$$K_{V_{ls}}^{(\overline{Y}),*}(f_\rho,\lambda) = \inf_{\alpha\in R^m}(\|f_\rho - f_\alpha\|_{2,\rho_X}^2 + \lambda m\|\alpha\|_2^2).$$

When the loss V is the hinge loss V_h, we have the following Theorem 1.3.

Theorem 1.3. *Let ρ be a (joint) finite nonnegative distribution on $Z, z = \{(x_i,w_i)\}_{i=1}^m$ be samples drawn randomly and independently according to $\rho(x,w)$. $K(x,y)$ is a given kernel on $X \times X$ and satisfies $k = \sup\limits_{(x,y)\in X\times X} |K(x,y)| < +\infty$. $\overline{Y} = \{y_1, y_2, \cdots, y_m\}$ is a given discrete set of X. α_z is the uniquely minimizer of scheme (4) for the hinge loss $V_h(t) = (1-t)_+$. Then, for any $0 < \delta < 1$, with confidence $1-\delta$, there holds*

$$\begin{aligned} & \Re(sgn(f_z)) - \Re(f_c) \\ \leq\ & \frac{4k^2(4log\frac{2}{\delta} + \sqrt{m})\ log\frac{2}{\delta}}{\lambda m} + K_{V_h}^{(\overline{Y})}(f_c,\lambda). \end{aligned} \tag{9}$$

After giving the explicit rates for $K_{V_{ls}}^{(\overline{Y}),*}(f_\rho,\lambda)$ and $K_{V_h}^{(\overline{Y})}(f_c,\lambda)$ respectively, we give the explicit excess estimates for (8) and (9) in Corollary 4.1 and Corollary 4.2, respectively.

2 The Sample Error

By [6, 37, 39] we know for a normalized loss $V(t)$ and a distribution ρ on Z there is a positive constant depending only upon V such that

$$\begin{aligned} & \Re(sgn(f)) - \Re(f_c) \\ \leq\ & \begin{cases} \mathcal{E}_{\rho,V}(f) - \mathcal{E}_{\rho,V}(f_c), \\ \text{if } V(t) = (1-t)_+; \\ \\ C_V\sqrt{\mathcal{E}_{\rho,V}(f) - \mathcal{E}_{\rho,V}(f^*)}, \\ \text{if } V''(0) \geq 0, \end{cases} \end{aligned} \tag{10}$$

if $\|f\|_\infty = \sup\limits_{x\in X} |f(x)| < +\infty$.

Eq.(10) shows that, to give the excess misclassification error (1), we need to bound the error

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

Since

$$\begin{aligned} & \mathcal{E}_{\rho,V}(f_{\alpha_z}) - \mathcal{E}_{\rho,V}(f^*) \\ \leq\ & |\mathcal{E}_{\rho,V}(f_{\alpha_z}) - \mathcal{E}_{\rho,V}(f_{\alpha^{(\rho)}})| \\ & + \mathcal{E}_{\rho,V}(f_{\alpha^{(\rho)}}) - \mathcal{E}_{\rho,V}(f^*), \end{aligned} \tag{11}$$

we need to estimate the sample error

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

and the approximation error

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

respectively. The approximation error can be estimated with approximation theory(see [26]). The sample error, nevertheless, is the main part of the whole estimate. The capacity-based approaches (see e.g.[1, 3, 5, 32, 33, 40]) and the capacity independent approaches (see e.g.[4, 8, 7, 9, 12, 21, 27]) are developed for these purposes.

In the present paper, we shall give an estimate for the sample error with convex analysis, the parallelogram identity and the large number law in Hilbert spaces. We first give the representations of the solutions of (4) and (5) with the derivatives of the loss, with which show the robustness of the solutions on the distributions. The sample error is then obtained.

Let $q \geq 1$ be a given positive integer and R^q be the q-dimensional Euclidean space. Then, for any $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_q) \in R^q$ and $\beta = (\beta_1, \beta_2, \cdots, \beta_q) \in R^q$ we can define in R^q the norm $\|\alpha\|_2 = (\sum\limits_{i=1}^{q} |\alpha_i|^2)^{\frac{1}{2}}$ and the inner product

$$(\alpha, \beta)_2 = \sum_{i=1}^{q} \alpha_i \beta_i.$$

It is well known that if $f(x)$ is a convex differentiable function on X, one has(see [18])

$$f(x') \geq f(x) + (\nabla f(x), x' - x)_2, \ x, x' \in R^m, \quad (12)$$

where $\nabla f(x)$ is the gradient of $f(x)$ at x.

Theorem 2.1. *If the conditions of Theorem 1.1 holds, then,for any $0 < \delta < 1$, with confidence $1 - \delta$, there is*

$$\begin{aligned} & |\mathcal{E}_{\rho,V}(f_{\alpha_z}) - \mathcal{E}_{\rho,V}(f_{\alpha^{(\rho)}})| \\ \leq & \ \frac{4k^2(4log\frac{2}{\delta} + \sqrt{m}) \ log\frac{2}{\delta}}{\lambda m} \\ & \times \|V'\|^2_{[-k\sqrt{\frac{V(0)}{\lambda}}, k\sqrt{\frac{V(0)}{\lambda}}]}. \end{aligned} \quad (13)$$

In particular, when $V = V_h$ is the hinge loss,there holds

$$\begin{aligned} & |\mathcal{E}_{\rho,V_h}(f_{\alpha_z}) - \mathcal{E}_{\rho,V_h}(f_{\alpha^{(\rho)}})| \\ \leq & \ \frac{4k^2(4log\frac{2}{\delta} + \sqrt{m}) \ log\frac{2}{\delta}}{\lambda m}. \end{aligned} \quad (14)$$

To show (13) we need some lemmas.

Lemma 2.1. *Let V be a given normalized loss, ρ be a distribution on Z. Then, there exists uniquely one minimizer of $\alpha^{(\rho)}$ of the problem (5) and*

$$\|\alpha^{(\rho)}\|_2 \leq \sqrt{\frac{V(0)}{m\lambda}}. \quad (15)$$

Proof. The uniqueness of the minimizer can be obtained by the fact that $V(t)$ is a convex function and $\lambda > 0$. Since $\alpha^{(\rho)}$ is the minimizer of (5), we have

$$\mathcal{E}_{\rho,V}(f_{\alpha^{(\rho)}}) + \lambda m \|\alpha^{(\rho)}\|_2^2 \leq \mathcal{E}_{\rho,V}(f_0) = V(0)$$

which gives (15).

We now give the representation of $\alpha^{(\rho)}$ with $V'(t)$.

Lemma 2.2. *Let V be a given normalized loss function, ρ be a distribution on Z.$\alpha^{(\rho)}$ is the solution of (5) for the given ρ. Define vector functions*

$$K_{\overline{Y}}(x) = (K(x, y_1), K(x, y_2), \cdots, K(x, y_m))^\top$$

for given discrete sets $\overline{Y} \subset Y$ and any $x \in X$. Then, there are the following results:

(i). Let ∇_α be the gradient of $V(wf_\alpha(x))$ about α. Then,

$$\begin{aligned} & \nabla_\alpha(V(wf_\alpha(x))) \\ = & \ K_{\overline{Y}}(x)^\top w V'(wf_\alpha(x)), \ (x, w) \in Z. \end{aligned} \quad (16)$$

(ii). The unique solution $\alpha^{(\rho)}$ of (5) has the following explicit expression

$$\alpha^{(\rho)} = -\frac{1}{2\lambda m} \int_Z K_{\overline{Y}}(x)^\top w V'(wf_{\alpha^{(\rho)}}(x)) d\rho, \quad (17)$$

where, for a vector function

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

and a function $\alpha(x)$, we define

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

and

$$\begin{aligned} & \int_Z f(x, w)\alpha(x) \ d\rho \\ & = (\int_Z f_1(x, w)\alpha(x) d\rho, \cdots, \int_Z f_m(x, w)\alpha(x) \ d\rho)^\top. \end{aligned}$$

Proof. (16) can be obtained by direct computations. By the definition of $\alpha^{(\rho)}$ we have

$$\begin{aligned} 0 &= \nabla_\alpha(\int_Z V(wf_\alpha(x)) d\rho + \lambda m \alpha^\top \alpha)|_{\alpha=\alpha^{(\rho)}} \\ &= \int_Z K_{\overline{Y}}(x)^\top w V'(wf_{\alpha^{(\rho)}}(x)) \ d\rho + 2\lambda \ m \ \alpha^{(\rho)}. \end{aligned}$$

(17) then holds.

Following Lemma 2.3 provides the connections among the solutions of (5) with respect to the distribution ρ.

Lemma 2.3. *Let V be a given normalized loss function, ρ and μ be distributions on $Z = X \times W$, $\alpha^{(\rho)}$ and $\alpha^{(\mu)}$ be the solutions of (5) for ρ and μ respectively. Then,*

$$\begin{aligned} & \|\alpha^{(\rho)} - \alpha^{(\mu)}\|_2 \\ \leq & \frac{1}{\lambda m} \| \int_Z K_{\overline{Y}}(x)^\top w V'(w f_{\alpha^{(\rho)}}(x)) d\rho \\ & - \int_Z K_{\overline{Y}}(x)^\top w V'(w f_{\alpha^{(\rho)}}(x)) d\mu \|_2. \quad (18) \end{aligned}$$

In particular, if μ is the empirical distribution $\mu_z(x, w)$ in (6), then, there is the following inequality

$$\begin{aligned} & \|\alpha^{(\rho)} - \alpha_z\|_2 \\ \leq & \frac{1}{\lambda m} \| \int_Z K_{\overline{Y}}(x)^\top w V'(w f_{\alpha^{(\rho)}}(x)) d\rho \\ - & \frac{1}{m} \sum_{i=1}^m K_{\overline{Y}}(x_i)^\top w_i V'(w_i f_{\alpha^{(\rho)}}(x_i)) \|_2. \quad (19) \end{aligned}$$

Proof of (18). By the convexity of $V(t)$ and (12) we have

$$\begin{aligned} & V(w f_{\alpha^{(\mu)}}(x)) - V(w f_{\alpha^{(\rho)}}(x)) \\ \geq & w V'(w f_{\alpha^{(\rho)}}(x)) K_{\overline{Y}}(x)^\top (\alpha^{(\mu)} - \alpha^{(\rho)}) \\ = & (\alpha^{(\mu)} - \alpha^{(\rho)}, w V'(w f_{\alpha^{(\rho)}}(x)) K_{\overline{Y}}(x)^\top)_2. \end{aligned}$$

It follows

$$\begin{aligned} & \int_Z V(w f_{\alpha^{(\mu)}}(x)) d\mu - \int_Z V(w f_{\alpha^{(\rho)}}(x)) d\mu \\ \geq & (\alpha^{(\mu)} - \alpha^{(\rho)}, \int_Z w V'(w f_{\alpha^{(\rho)}}(x)) K_{\overline{Y}}(x)^\top d\mu)_2. \end{aligned}$$

Since $\alpha^{(\mu)}, \alpha^{(\rho)} \in R^m$, then, the reformed parallelogram equality gives

$$\begin{aligned} \|\alpha^{(\mu)}\|_2^2 - \|\alpha^{(\rho)}\|_2^2 = & 2(\alpha^{(\mu)} - \alpha^{(\rho)}, \alpha^{(\rho)})_2 \\ & + \|\alpha^{(\rho)} - \alpha^{(\mu)}\|_2^2. \quad (20) \end{aligned}$$

It follows by (20) that

$$\begin{aligned} & (\mathcal{E}_{\mu,V}(f_{\alpha^{(\mu)}}) + \lambda m \|\alpha^{(\mu)}\|_2^2) \\ & - (\mathcal{E}_{\mu,V}(f_{\alpha^{(\rho)}}) + \lambda m \|\alpha^{(\rho)}\|_2^2) \\ \geq & (\alpha^{(\mu)} - \alpha^{(\rho)}, \int_Z w V'(w f_{\alpha^{(\rho)}}(x)) K_{\overline{Y}}(x)^\top d\mu)_2 \\ & + 2\lambda m (\alpha^{(\mu)} - \alpha^{(\rho)}, \alpha^{(\rho)})_2 \\ & + \lambda m \|\alpha^{(\rho)} - \alpha^{(\mu)}\|_2^2. \\ \geq & (\alpha^{(\mu)} - \alpha^{(\rho)}, \int_Z w V'(w f_{\alpha^{(\rho)}}(x)) K_{\overline{Y}}(x)^\top d\mu \\ & - \int_Z w V'(w f_{\alpha^{(\rho)}}(x)) \, K_{\overline{Y}}(x)^\top d\rho)_2 \\ & + \lambda m \|\alpha^{(\rho)} - \alpha^{(\mu)}\|_2^2. \end{aligned}$$

Since

$$\begin{aligned} & (\mathcal{E}_{\mu,V}(f_{\alpha^{(\mu)}}) + \lambda m \|\alpha^{(\mu)}\|_2^2) - (\mathcal{E}_{\mu,V}(f_{\alpha^{(\rho)}}) \\ & + \lambda m \|\alpha^{(\rho)}\|_2^2) \leq 0, \end{aligned}$$

we have by the Cauchy inequality that

$$\begin{aligned} & \lambda m \|\alpha^{(\rho)} - \alpha^{(\mu)}\|_2^2 \\ \leq & (\alpha^{(\rho)} - \alpha^{(\mu)}, \int_Z w V'(w f_{\alpha^{(\rho)}}(x)) K_{\overline{Y}}(x)^\top \, d\mu \\ & - \int_Z w V'(w f_{\alpha^{(\rho)}}(x)) \, K_{\overline{Y}}(x)^\top d\rho)_2 \\ \leq & \|\alpha^{(\rho)} - \alpha^{(\mu)}\|_2 \times \| \int_Z w V'(w f_{\alpha^{(\rho)}}(x)) K_{\overline{Y}}(x)^\top \, d\mu \\ & - \int_Z w V'(w f_{\alpha^{(\rho)}}(x)) K_{\overline{Y}}(x)^\top \, d\rho \|_2. \end{aligned}$$

Consequently, (18) holds. (19) can be followed by taking $\mu = \mu_z$.

Following large number law will play a vital role in proving Theorem 2.1.

Lemma 2.4. (See [21]). *Let 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. Denote $\sigma^2(\xi) = E(\|\xi\|_H^2)$ and let $\{\xi_i\}_{i=1}^m$ be independent random drawers of ρ. Then, for any $0 < \delta < 1$, with confidence $1 - \delta$, there holds*

$$\begin{aligned} & \| \frac{1}{m} \sum_{i=1}^m (\xi_i - E(\xi_i)) \|_H \\ \leq & \frac{2\widetilde{M} log(2/\delta)}{m} + \sqrt{\frac{2\sigma^2(\xi) log(2/\delta)}{m}}. \quad (21) \end{aligned}$$

Proof of (13). By the definition of $\alpha^{(\rho)}$ we have

$$\begin{aligned} & |\mathcal{E}_{\rho,V}(f_{\alpha_z}) - \mathcal{E}_{\rho,V}(f_{\alpha^{(\rho)}})| \\ \leq & (\mathcal{E}_{\rho,V}(f_{\alpha_z}) + \lambda m \|\alpha_z\|_2^2) - (\mathcal{E}_{\rho,V}(f_{\alpha^{(\rho)}}) \\ & + \lambda m \|\alpha^{(\rho)}\|_2^2) + \lambda m | \|\alpha_z\|_2^2 - \|\alpha^{(\rho)}\|_2^2 | \\ = & A + \lambda m | \|\alpha_z\|_2^2 - \|\alpha^{(\rho)}\|_2^2 |, \quad (22) \end{aligned}$$

where

$$\begin{aligned} A = & \mathcal{E}_{\rho,V}(f_{\alpha_z}) - \mathcal{E}_{\rho,V}(f_{\alpha^{(\rho)}}) \\ & + \lambda m (\|\alpha_z\|_2^2) - \|\alpha^{(\rho)}\|_2^2). \end{aligned}$$

Rewrite (12) by

$$f(x)-f(x')\le(\nabla f(x),x-x')_2,\qquad x,x'\in R^m.$$

Then,

$$\begin{aligned}&\mathcal{E}_{\rho,V}(f_{\alpha_z})-\mathcal{E}_{\rho,V}(f_{\alpha^{(\rho)}})\\ =\ &\int_Z V(wf_{\alpha_z}(x))d\rho-\int_Z V(wf_{\alpha^{(\rho)}}(x))d\rho\\ \le\ &(\int_Z \nabla_\alpha V(wf_\alpha(x))|_{\alpha=\alpha_z}\,d\rho,\alpha_z-\alpha^{(\rho)})_2.\end{aligned}$$

On the other hand, by (20) we have

$$\begin{aligned}&\|\alpha_z\|_2^2-\|\alpha^{(\rho)}\|_2^2\\ =\ &2(\alpha_z-\alpha^{(\rho)},\alpha_z)_2-\|\alpha^{(\rho)}-\alpha^{(\mu)}\|_2^2.\end{aligned}\quad(23)$$

It follows

$$\begin{aligned}A\ \le\ &(\int_Z \nabla_\alpha V(wf_\alpha(x))|_{\alpha=\alpha_z}\,d\rho,\alpha_z-\alpha^{(\rho)})_2\\ &+2\lambda m(\alpha_z,\alpha_z-\alpha^{(\rho)})_2\\ &-\lambda m\|\alpha_z-\alpha^{(\rho)}\|_2^2\\ =\ &(\int_Z wV'(wf_{\alpha_z}(x))K_{\overline{Y}}^\top(x)\,d\rho+2\lambda m\alpha_z,\\ &\alpha_z-\alpha^{(\rho)})_2-\lambda m\|\alpha_z-\alpha^{(\rho)}\|_2^2\\ \le\ &(\int_Z wV'(wf_{\alpha_z}(x))K_{\overline{Y}}^\top(x)\,d\rho\\ &-\frac{1}{m}\sum_{i=1}^m w_iV'(w_if_{\alpha_z}(x_i))K_{\overline{Y}}^\top(x_i),\\ &\alpha_z-\alpha^{(\rho)})_2-\lambda m\|\alpha^{(\rho)}-\alpha_z\|_2^2.\end{aligned}\quad(24)$$

(23),(22) and (24) gives

$$\begin{aligned}&|\mathcal{E}_{\rho,V}(f_{\alpha_z})-\mathcal{E}_{\rho,V}(f_{\alpha^{(\rho)}})|\\ \le\ &(\int_Z wV'(wf_{\alpha_z}(x))K_{\overline{Y}}^\top(x)\,d\rho\\ &-\frac{1}{m}\sum_{i=1}^m w_iV'(w_if_{\alpha_z}(x_i))K_{\overline{Y}}^\top(x_i),\alpha_z-\alpha^{(\rho)})_2\\ &-\lambda m\|\alpha^{(\rho)}-\alpha_z\|_2^2+\lambda m|\|\alpha_z\|_2^2-\|\alpha^{(\rho)}\|_2^2|\\ \le\ &(\int_Z wV'(wf_{\alpha_z}(x))K_{\overline{Y}}^\top(x)\,d\rho\\ &-\frac{1}{m}\sum_{i=1}^m w_iV'(w_if_{\alpha_z}(x_i))K_{\overline{Y}}^\top(x_i),\alpha_z-\alpha^{(\rho)})_2\\ &-\lambda m\|\alpha^{(\rho)}-\alpha_z\|_2^2+2\lambda m|(\alpha_z,\alpha^{(\rho)}-\alpha_z)_2|\\ &+\lambda m\|\alpha^{(\rho)}-\alpha_z\|_2^2\\ \le\ &(\int_Z wV'(wf_{\alpha_z}(x))K_{\overline{Y}}(x)^\top\,d\rho\\ &-\frac{1}{m}\sum_{i=1}^m w_iV'(w_if_{\alpha_z}(x_i))K_{\overline{Y}}^\top(x_i),\alpha_z-\alpha^{(\rho)})_2\\ &+\ 2\lambda m\|\alpha_z\|_2\times\|\alpha^{(\rho)}-\alpha_z\|_2\\ \le\ &(\|\int_Z wV'(wf_{\alpha_z}(x))K_{\overline{Y}}(x)^\top\,d\rho\\ &-\frac{1}{m}\sum_{i=1}^m w_iV'(w_if_{\alpha_z}(x_i))K_{\overline{Y}}(x_i)^\top\|_2\\ &+\|\frac{1}{m}\sum_{i=1}^m w_iV'(w_if_{\alpha_z}(x_i))K_{\overline{Y}}(x_i)^\top\|_2)\\ &\times\|\alpha_z-\alpha^{(\rho)}\|_2.\end{aligned}\quad(25)$$

Since (15), we have

$$\|f_{\alpha^{(\rho)}}\|_\infty\le\sqrt{m}k\times\|\alpha^{(\rho)}\|_2\le k\sqrt{\frac{V(0)}{\lambda}},$$

and

$$\|f_{\alpha_z}\|_\infty\le\sqrt{m}k\times\|\alpha_z\|_2\le k\sqrt{\frac{V(0)}{\lambda}}.$$

Then, for any $x\in X$ there holds

$$\begin{aligned}&\|wV'(wf_{\alpha_z}(x))\ K_{\overline{Y}}^\top(x)\|_2\\ \le\ &k\sqrt{m}\times\|V'(wf_{\alpha_z})\|_\infty\\ \le\ &k\sqrt{m}\times\|V'\|_{[-k\sqrt{\frac{V(0)}{\lambda}},k\sqrt{\frac{V(0)}{\lambda}}]},\end{aligned}\quad(26)$$

and

$$\begin{aligned}&\|wV'(wf_{\alpha^{(\rho)}}(x))K_{\overline{Y}}^\top(x)\|_2\\ \le\ &k\sqrt{m}\times\|V'(wf_{\alpha^{(\rho)}})\|_\infty\\ \le\ &k\sqrt{m}\times\|V'\|_{[-k\sqrt{\frac{V(0)}{\lambda}},k\sqrt{\frac{V(0)}{\lambda}}]}.\end{aligned}\quad(27)$$

Therefore,by (26)

$$\begin{aligned}&\|\int_Z wV'(wf_{\alpha_z}(x))K_{\overline{Y}}(x)^\top\,d\rho\\ &-\frac{1}{m}\sum_{i=1}^m\ w_iV'(w_if_{\alpha_z}(x_i))\ K_{\overline{Y}}(x_i)^\top\|_2\\ \le\ &\sup_{\|h(x,w)\|_2\le k\sqrt{m}\times\|V'\|_{[-k\sqrt{\frac{V(0)}{\lambda}},k\sqrt{\frac{V(0)}{\lambda}}]}}\\ &\|\int_Z h(x,w)\,d\rho-\frac{1}{m}\sum_{i=1}^m h(x_i,w_i)\|_2\end{aligned}\quad(28)$$

and by (27) we have

$$\begin{aligned}&\|\int_Z wV'(wf_{\alpha^{(\rho)}}(x))K_{\overline{Y}}(x)^\top\,d\rho\\ &-\frac{1}{m}\sum_{i=1}^m\ w_iV'(w_if_{\alpha^{(\rho)}}(x_i))\ K_{\overline{Y}}(x_i)^\top\|_2\end{aligned}$$

$$\leq \sup_{\|h(x,w)\|_2 \leq k\sqrt{m}\times\|V'\|_{[-k\sqrt{\frac{V(0)}{\lambda}},k\sqrt{\frac{V(0)}{\lambda}}]}} \times\| \int_Z h(x,w)\ d\rho - \frac{1}{m}\sum_{i=1}^m h(x_i,w_i)\|_2. \quad (29)$$

On the other hand, by (26) we have

$$\|\frac{1}{m}\sum_{i=1}^m w_i V'(w_i f_{\alpha_z}(x_i))K_{\overline{Y}}(x_i)^\top\|_2 \leq k\sqrt{m}\times\|V'\|_{[-k\sqrt{\frac{V(0)}{\lambda}},k\sqrt{\frac{V(0)}{\lambda}}]}. \quad (30)$$

(19), (29), (30),(21) and (25) give

$$\begin{aligned} &|\mathcal{E}_{\rho,V}(f_{\alpha_z}) - \mathcal{E}_{\rho,V}(f_{\alpha^{(\rho)}})| \\ \leq\ & \frac{1}{\lambda m}\times\Big(\sup_{\|h(x,w)\|_2 \leq k\sqrt{m}\times\|V'\|_{[-k\sqrt{\frac{V(0)}{\lambda}},k\sqrt{\frac{V(0)}{\lambda}}]}} \\ & \times\| \int_Z h(x,w)\ d\rho - \frac{1}{m}\sum_{i=1}^m h(x_i,w_i)\|_2 \\ & +k\sqrt{m}\times\|V'\|_{[-k\sqrt{\frac{V(0)}{\lambda}},k\sqrt{\frac{V(0)}{\lambda}}]}\Big) \\ & \times \sup_{\|h(x,w)\|_2 \leq k\sqrt{m}\times\|V'\|_{[-k\sqrt{\frac{V(0)}{\lambda}},k\sqrt{\frac{V(0)}{\lambda}}]}} \\ & \times\| \int_Z h(x,w)\ d\rho - \frac{1}{m}\sum_{i=1}^m h(x_i,w_i)\|_2. \end{aligned} \quad (31)$$

By (21) we know, with confidence $1-\delta$, there holds

$$\begin{aligned} & \sup_{\|h(x,w)\|_2 \leq k\sqrt{m}\times\|V'\|_{[-k\sqrt{\frac{V(0)}{\lambda}},k\sqrt{\frac{V(0)}{\lambda}}]}} \\ & \times\| \int_Z h(x,w)\ d\rho - \frac{1}{m}\sum_{i=1}^m h(x_i,w_i)\|_2 \\ \leq\ & k\sqrt{m}\times(\frac{2log(2/\delta)}{m}+\sqrt{\frac{2log(2/\delta)}{m}}) \\ & \times\|V'\|_{[-k\sqrt{\frac{V(0)}{\lambda}},k\sqrt{\frac{V(0)}{\lambda}}]} \\ \leq\ & 4klog(2/\delta)\times\|V'\|_{[-k\sqrt{\frac{V(0)}{\lambda}},k\sqrt{\frac{V(0)}{\lambda}}]}. \end{aligned} \quad (32)$$

(32) and (31) give

$$\begin{aligned} &|\mathcal{E}_{\rho,V}(f_{\alpha_z}) - \mathcal{E}_{\rho,V}(f_{\alpha^{(\rho)}})| \\ \leq\ & \frac{1}{\lambda m}\times\Big(4klog(2/\delta)\times\|V'\|_{[-k\sqrt{\frac{V(0)}{\lambda}},k\sqrt{\frac{V(0)}{\lambda}}]} \\ & +k\sqrt{m}\times\|V'\|_{[-k\sqrt{\frac{V(0)}{\lambda}},k\sqrt{\frac{V(0)}{\lambda}}]}\Big) \\ & \times 4klog(2/\delta)\times\|V'\|_{[-k\sqrt{\frac{V(0)}{\lambda}},k\sqrt{\frac{V(0)}{\lambda}}]}. \end{aligned}$$

(13) then holds.

Proof of (14). Simple computations give $V_h'(t) = -1$ if $t \leq 1$ and 0 if $t > 0$. We then have $\|V'\|_{[-k\sqrt{\frac{V(0)}{\lambda}},k\sqrt{\frac{V(0)}{\lambda}}]} = 1$. (14) follows from (13).

3 The Approximation Error

We now show the approximation error.

Theorem 3.1. *Let V be a given normalized loss function, ρ be a distribution on $Z = X \times W$, $\alpha^{(\rho)}$ be the solutions of (5) for ρ. Then,*

$$\mathcal{E}_{\rho,V}(f_{\alpha^{(\rho)}}) - \mathcal{E}_{\rho,V}(f^*) \leq K_V^{(\overline{Y})}(f^*,\lambda). \quad (33)$$

(33) shows that the approximation error is bounded by the K-functional $K_V^{(\overline{Y})}(f^*,\lambda)$.

Proof. By the definitions of $\alpha^{(\rho)}$ and f^* we have

$$\begin{aligned} 0 &\leq \mathcal{E}_{\rho,V}(f_{\alpha^{(\alpha)}}) - \mathcal{E}_{\rho,V}(f^*) \\ &\leq \mathcal{E}_{\rho,V}(f_{\alpha^{(\rho)}}) - \mathcal{E}_{\rho,V}(f^*) + \lambda m\|\alpha^{(\alpha)}\|_2^2 \\ &= \inf_{\alpha\in R^m}(\mathcal{E}_{\rho,V}(f_\alpha) + \lambda m\|\alpha\|_2^2 - \mathcal{E}_{\rho,V}(f^*)) \\ &= K_V^{(\overline{Y})}(f^*,\lambda). \end{aligned}$$

(33) then holds. If V_h is the hinge loss, then, $f^* = f_c$. In this case, we may give an estimate for $K_{V_h}^{(\overline{Y})}(f^*,\lambda)$.

Theorem 3.2. *Let $\varphi \in L^2(\rho_X)$ satisfy*

$$f^*(x) = \int_X K(x,y)\varphi(y)d\rho_X(y), \qquad x \in X. \quad (34)$$

Then, there is a discrete $\overline{Y} \subset X$ such that

$$K_{V_h}^{(\overline{Y})}(f^*,\lambda) \leq \sqrt{\frac{A-\|f_\rho\|_{2,\rho_X}^2}{m}} + \lambda\|\varphi\|_{2,\rho_X}^2. \quad (35)$$

To show (35) we need a lemma.

Lemma 3.1.(see [24]) *Let f^* satisfy (34). Then, there is a discrete set $\overline{Y} \subset X$ and an $\alpha^* \in R^m$ such that*

$$\|f^* - f_{\alpha^*}\|_{2,\rho_X} \leq \sqrt{\frac{A-\|f_\rho\|_{2,\rho_X}^2}{m}} \quad (36)$$

and $m\|\alpha^*\|_2^2 \le \|\varphi\|_{2,\rho_X}^2$.

Proof of Theorem 3.2. Since V_h is a Lipschitz function with Lipschitz constant 1, i.e.,

$$|V_h(t) - V_h(t')| \le |t - t'|, \qquad t, t' \in R,$$

we have by Lemma 3.1 that there is an $\alpha^* \in R^m$ and a discrete set $\overline{Y} \subset X$ that

$$\begin{aligned}
& K_{V_h}^{(\overline{Y})}(f_c, \lambda) \\
= & \inf_{\alpha \in R^m} (|\mathcal{E}_{\rho,V_h}(f_\alpha) - \mathcal{E}_{\rho,V_h}(f_c)| + \lambda m \|\alpha\|_2^2) \\
\le & \inf_{\alpha \in R^m} (\int_Z |V_h(wf_\alpha(x)) - V_h(wf_c(x))| d\rho \\
& + \lambda m \|\alpha\|_2^2) \\
\le & \inf_{\alpha \in R^m} (\int_X |f_\alpha(x) - f_c(x)| d\rho + \lambda m \|\alpha\|_2^2) \\
\le & \inf_{\alpha \in R^m} (\|f_\alpha - f_c\|_{2,\rho_X} + \lambda m \|\alpha\|_2^2) \\
\le & \|f_c - f_{\alpha^*}\|_{2,\rho_X} + \lambda m \|\alpha^*\|_2^2 \\
\le & \sqrt{\frac{A - \|f_c\|_{2,\rho_X}^2}{m}} + \lambda \|\varphi\|_{2,\rho_X}^2.
\end{aligned}$$

4 Proof of the Results

Proof of Theorem 1.1. By (10),(11),(13)and (33) we have (7).

Proof of Theorem 1.3. By (10),(11),(14) and (33) we have (9).

Proof of Theorem 1.2. Since $V_{ls}''(t) = 2 > 0$, we know by (10) that

$$\begin{aligned}
& \Re(sgn(f_z)) - \Re(f_c) \\
\le & C_V \sqrt{\mathcal{E}_{\rho,V_{ls}}(f_z) - \mathcal{E}_{\rho,V_{ls}}(f_\rho)}. \qquad (37)
\end{aligned}$$

Since

$$\begin{aligned}
\mathcal{E}_{\rho,V_{ls}}(f_z) &= \int_Z (1 - wf_z(x))^2 d\rho \\
&= \int_Z (w - f_z(x))^2 d\rho,
\end{aligned}$$

we have by [10] the equality

$$\mathcal{E}_{\rho,V_{ls}}(f_z) - \mathcal{E}_{\rho,V_{ls}}(f_\rho) = \|f_z - f_\rho\|_{2,\rho}^2,$$

which and (37) give

$$\Re(sgn(f_z)) - \Re(f_c) \le C_V \|f_z - f_\rho\|_{2,\rho}. \qquad (38)$$

On the other hand, by Theorem 2 of [25] we know

$$\begin{aligned}
\|f_z - f_\rho\|_{2,\rho} &\le \frac{6k^2 \log \frac{2}{\delta}}{\lambda\sqrt{m}} \\
&+ \sqrt{K_{V_{ls}}^{(\overline{Y}),*}(f_\rho, \lambda)}. \qquad (39)
\end{aligned}$$

(38) and (39) give (8).

Corollary 4.1. *Under the conditions of Theorem 1.3, if f_c satisfies (34), then, there is a discrete set $\overline{Y} \subset X$ such that, for any $0 < \delta < 1$, with confidence $1 - \delta$, there holds*

$$\begin{aligned}
& \Re(sgn(f_z)) - \Re(f_c) \\
\le & \frac{4k^2(4log\frac{2}{\delta} + \sqrt{m})\ log\frac{2}{\delta}}{\lambda m} \\
& + \sqrt{\frac{A - \|f_c\|_{2,\rho_X}^2}{m}} + \lambda \|\varphi\|_{2,\rho_X}^2. \qquad (40)
\end{aligned}$$

Proof. (40) can be obtained by (9) and (35).

Corollary 4.2. *Under the conditions of Theorem 1.2, if f_ρ satisfies (34), then, there is a constant $C > 0$ such that, for any $0 < \delta < 1$, with confidence $1 - \delta$, there holds*

$$\begin{aligned}
& \Re(sgn(f_z)) - \Re(f_c) \\
\le & \frac{6k^2 \log \frac{2}{\delta}}{\lambda\sqrt{m}} + \sqrt{\frac{A - \|f_c\|_{2,\rho_X}^2}{m}} \\
& + \sqrt{\lambda} \|\varphi\|_{2,\rho_X}. \qquad (41)
\end{aligned}$$

Proof. (41) can be obtained by Lemma 3.1 and (8).

Acknowledgements: The research is supported by NSF of China (grant No. 10871226, 11001247) and the NSF of Zhejiang province (grant No. Y6100096).

References:

[1] N. Alon, S. Ben-David, N. Cesa-Bianchi, D. Haussler, Scale-sensitive dimension uniform convergence and learnability, *J. Assoc. Comput. Mach.*, 44,1997, pp.615-631.

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

[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] P. L. Bartlett, M. I. Jordan, J. D. McAuliffe, Convex,classification,and risk bounds, *J. Amer. Statist. Assoc.*, 101, 2006, pp. 138-156.

[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] D. R. Chen, Q. Wu, Y. Ying, D. X. Zhou, Support vector machine soft margin classifiers: error analysis, *J. Mach. Learn. Res.*, 5, 2004, pp. 1143-1175.

[7] D. R. Chen, Y. Xu, Minimax optimal rates of convergence for multicategory classification, *Acta Math. Sinica, English Series*, 23(8), 2007, pp.1419-1426.

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

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

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

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

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

[13] L. Devroye,L. Györfi, G. Lugosi, *A Probability Theory of Pattern Recognition*, Springer-Verlag, New York, 1997.

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

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

[16] W. Fu, Penalized regressions: the bridge versus the lasso, *J. Comput. Graph. Statist.*, 7(3), 1998, pp.397-416.

[17] E. Greenshtein, Best subset selection, persistence in high-dimensional ststistical learning and optimization under l_1- constraint, *Ann. Statist.* 34(5), 2006, pp. 2367-3386.

[18] J.-B., Hiriart-Urruty, C. Lemaréchas, *Fundamental of Convex Analysis*, Springer-verlag, Berlin, 2001.

[19] V. Koltchinskii, Sparisity in penalized empirical risk minimization, *Annales de l'lnstitut Henri Poincaré B, Probab. Statist.*45(1), 2009, pp. 7-57.

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

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

[22] S. Smale, D. X. Zhou, Shannon sampling and function reconstruction from point values, *Bull.(New Series) Amer. Math. Soc.*,41(3), 2004, pp. 279-305

[23] B. H. Sheng, D. H. Xiang, The Convergence Rate for a K-functional in Learning Theory, *J. Inequalities and Applications*,Volume 2010, Article ID 249507, 18 pages doi: 10. 1155/ 2010/ 249507

[24] B. H. Sheng, P. X. Ye, J. L. Wang, Learning rates for least square regressions with coefficient regularization, *Acta Mathematica Sinica*, English Series (accepted).

[25] 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.

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

[27] I. Steinwart, Sparseness of support vector machines, *J. Mach. Learn. Res.*, 4,2003, pp. 1071-1105.

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

[29] B. Tarigan, S. A. Van de Geer, Classifiers of support vector machine type with l_1- complexity regularization, *Bernoulli*,12(6),2006, pp. 1045-1076.

[30] H. Z. Tong, D. R. Chen, L. Z. Peng, Learning rates for regularized classifiers using multivariate polynomial kernels, *J. Complexity*, 24,2008, pp. 619-631.

[31] S. A. Van De Geer, High-dimensional generalized linear models and the lasso, *Ann. Statist.*, 36(2),2008, pp. 614-645.

[32] S. A. Van De Geer, *Applications of Empirical Process Theory*, Cambridge University Press, Cambridge, 2000.

[33] R. C. Williamson, A. J. Smola, and B.Scholkopf, Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators, *IEEE Trans. Inform. Theory*, 47, 2001, pp.2516-2532.

[34] Q. Wu, D. X. Zhou, SVM soft margin classifiers: linear programming versus quadratic programming, *Neural Comput.*, 17,2005, pp.1160-1187.

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

[36] Q. Wu, D. X.Zhou, Analysis of support vector machine classification, *J. Comput. Anal. Appl.* 8,2006, pp. 99-119.

[37] Q. Wu, Y. Ying, D. X. Zhou, Multi-kernel regularized classifiers, *J. Complexity*, 23,2007, pp. 108-134.

[38] Q. W. Xiao, D. X. Zhou, Learning by nonsymmetric kernels with data dependent spaces and l^1-regularizer, *Taiwanese J. Math.*, 14(5), 2010, pp. 1821-1836

[39] T. Zhang, Statistical behavior and consistency of classificattion methods based on convex risk minimization, *The Ann. Statist*. 32(1), 2004, pp. 56-134.

[40] D. X. Zhou, Capacity of reproducing kernel spaces in learning theory, *IEEE Trans. Inform. Theory*, 49, 2003, pp. 1743-1752.