Learning Rates of Support Vector Machine Classifiers with Data Dependent Hypothesis Spaces

Bao-Huai Sheng
Department of Mathematics, Shaoxing College of Arts and Sciences Shaoxing, Zhejiang 312000, China
e-mail: shengbaohuai@163.com

Pei-Xin Ye
School of Mathematics and LPMC, Nankai University, Tianjin 300071, China
e-mail: yepx@nankai.edu.cn

Abstract—We study the error performances of \( p \)-norm Support Vector Machine classifiers based on reproducing kernel Hilbert spaces. We focus on two category problem and choose the data-dependent polynomial kernels as the Mercer kernel to improve the approximation error. We also provide the standard estimation of the sample error, and derive the explicit learning rate.

Index Terms—Support vector machine classification; Learning rate; Reproducing kernel Hilbert spaces; Cesaro means.

I. INTRODUCTION AND RESULTS

Support vector machine classification [1]-[7], [9]-[25] has a foundation in the framework of statistical learning theory and classical regularization theory for function approximation. It is one of the most important topics in the field of machine learning. It has been applied successfully to various practical problems in science, engineering and many other related fields. The goal of classification is to construct a classifier which can predict the unknown class of an observation with small misclassification error. This problem has been studied widely and many important algorithms have been developed (Ref. [3]-[8]).

Let \( X = [-1,1] \), \( Y = \{-1,1\} \). A binary classifier \( f : X \rightarrow Y \) divides the input space \( X \) into two classes.

Let \( \rho \) be an unknown probability distribution on \( X \times Y \) and \( (X,Y) \) be the corresponding random variable. The misclassification error for a classifier \( f : X \rightarrow Y \) is defined to be the probability of the event \( \{f(X) \neq Y\} \)

\[ \mathcal{R}(f) := \text{Prob}\{f(X) \neq Y\} = \int_X P(Y \neq f(x)|x)d\rho_x(x), \]

where \( \rho_x \) is the marginal distribution on \( X \) and \( \rho(y|x) \) is the conditional probability measure at \( x \) induced by \( \rho \). The distribution \( \rho \) is known only through a set of samples \( z := \{z_i\}_{i=1}^n = \{(x_i,y_i)\}_{i=1}^n \in \mathbb{Z}^n \) independently drawn according to \( \rho \). It is known from [9] the classifier which minimizes the misclassification error is the Bayes ruler \( f_\rho := \text{sgn}(f_\rho) \), where

\[ f_\rho(x) = \int yd\rho(y|x) = P(y = 1|x) - P(y = -1|x), \quad x \in X. \]

Denote the \( p \)-norm hinge loss function as

\[ V(yf(x)) := (1-yf(x))^p, \]

where \( (\cdot,\cdot) \) is the indicator function. \( V(yf(x)) \) measures the cost paid by replacing the true \( y \) with the estimate \( f(x) \). The corresponding \( V \)-risk is

\[ \mathcal{E}(f) := \int_{-1}^{1} V(yf(x))d\rho = EV(yf(x)). \]

Let \( f^*_\rho : X \rightarrow R \) be a measurable function minimizing the expected risk \( f^*_\rho := \arg \min \mathcal{E}(f) \), where the minimum is taken over all measurable functions. According to [8], we may always choose a \( f^*_\rho \) such that \( f^*_\rho(x) \in [-1,1] \) for each \( x \in X \). Since the expected risk involving the unknown distribution \( \rho \) is not computable, its discretization is used instead which is computable in terms of the sample \( z \), is defined as

\[ \mathcal{E}_z(f) := \frac{1}{m} \sum_{i=1}^{m} V(y,z_i)f(z_i) = \frac{1}{m} \sum_{i=1}^{m} (1 - y_i f(z_i))^p. \]

Regularized learning schemes are implemented by minimizing a penalized version of the empirical error over a set of functions, called a hypothesis space. Then the regularized classifier generated for a sample \( z \in \mathbb{Z}^n \) is defined as \( \text{sgn}(f_{\lambda,z}) \), where \( f_{\lambda,z} \) is a minimizer of the following well-known Tikhonov regularization scheme

\[ f_{\lambda,z} := \arg \min_{f \in \Omega} \left\{ \frac{1}{m} \sum_{i=1}^{m} V(y,z_i)f(z_i) + \lambda \Omega(f) \right\}. \]

Here \( \lambda \) is called regularization parameter, it depends on \( m \) and usually \( \lambda(m) \rightarrow 0 \) as \( m \) becomes large. This
kind of scheme is very popular in many areas of the applications and theory of machine learning.

In this paper, we take the hypothesis space \( H \) to be the reproducing kernel spaces reproducing by polynomials on \([-1,1] \times [-1,1]\) and a given data in \([-1,1]\). We want to estimate excess misclassification error \( R(\text{sgm}(f_{\lambda k})) - R(f_{\lambda}) \). Many investigations of this topics have been done when the kernel \( K(x, y) \) is the Gaussian kernel \( \kappa_\sigma(x, t) = \exp\left[-\frac{k - t}{2\sigma^2}\right] \) (Ref. e.g.\([5],[8]\)). Now let us turn to other kernels, when the kernel is a polynomial kernel, the excess misclassification error estimate may become simpler since the polynomial class has many advantages in estimating the covering number, which is needed in presenting the error estimate. In fact, the excess misclassification error estimate for the polynomial kernel is a field investigated by many mathematicians. Among others, \([5]\) gave a quantitative estimate for the convergence rate in the univariate case \( X = [0,1] \) with the Bernstein- Durrmeyer polynomials operators, \([4]\) gave the estimate in the case \( X \subset \mathbb{R} \) being a simplex. It is known that the best error of the Bernstein operator is \( O\left(\frac{1}{\sqrt{m}}\right) \). Thus we introduce the generalized Vallee Poussin means of orthonormal algebraic polynomials which can approximate the target function with \( O\left(\frac{1}{m^\beta}\right) \) for some \( \alpha > 0 \), and the orthogonal algebraic polynomials may exist for any positive measure \( \rho_k \) (Ref.e.g.\([12],[13]\)). These properties is much better than the Bernstein operator and it enable us to estimate the excess misclassification error by generalized Vallee Poussin means. So we may yield better approximation error and therefore the learning error can be improved.

Note that now our hypothesis space \( \mathcal{H} \) depends on the marginal distribution \( \lambda = \int x \rho(x) dx \) with \( \rho(x) \) being Jacobi weights on \([-1, 1]\). It is known that the best properties is much better than the Bernstein polynomials \( \mathcal{H} \). We want to write \( \lambda \) such that \( \lambda = \sum_{i=0}^{n} \eta_i \eta_i(\lambda) \rho_i(\lambda) \) and assume the marginal distribution \( \lambda \) being a simplex. Therefore, we can express the generating function \( \eta(x) \) as a linear combination of \( \eta_i \). These properties is much better than the Bernstein polynomials. We have \( \lambda \) such that \( \lambda = \sum_{i=0}^{n} \eta_i \eta_i(\lambda) \rho_i(\lambda) \), \( \forall x \in [-1,1] \).

Throughout the paper, we shall write \( A = O(B) \) if there exists a constant \( C > 0 \) such that \( A \leq CB \). We write \( A \sim B \) if \( A = O(B) \) and \( B = O(A) \). Let \( X = [-1,1] \) and assume the marginal distribution \( \rho_k(x) = w(x) dx \) with \( w(x) = (1-x)^{\alpha_1}(1+x)^{\alpha_2} \), \( \alpha_i > -1 \), \( \alpha_i + \alpha_j > -1 \), being Jacobi weights on \([-1,1]\) with finite moments; i.e., \( \int_{-1}^{1} x^j w(x) dx < \infty \), \( k \in \mathbb{N}_0 := 0, 1, 2, \ldots \). It is known from \([14]\) that for all \( k \in \mathbb{N}_0 \) there exists a unique polynomial

\[
p_k(x) = \gamma_k(w)x^k + \cdots, \quad \gamma_k(w) > 0
\]

such that

\[
\int_{-1}^{1} p_k(x)p_k(x)w(x) dx = \delta_{k,k}.
\]

Then we denote

\[
a_k(f) = \int_{-1}^{1} f(x)p_k(x)w(x) dx, \quad k \in \mathbb{N}_0.
\]

Moreover, there uniquely exists Lagrange polynomial interpolating operator \( L_k(x) \) of order \( n - 1 \), such that

\[
L_k(x) = y_k, \quad k = 1, 2, \ldots, n,
\]

for any real numbers \( y_k \), \( k = 1, 2, \ldots, n \).

Let \( p \geq 1 \), \( L_p(dw) \) be the class of all measurable real functions \( f \) for which

\[
\|f\|_{L_p(dw)} = \left(\int_{-1}^{1} |f(x)|^p w(x) dx \right)^{\frac{1}{p}} < \infty.
\]

For \( \delta > \max(\alpha_1 + \frac{1}{2}, \alpha_2 + \frac{1}{2}) \) the Cesaro means \( C(\delta) \) of \( f \) is defined by

\[
\sigma_n^\delta(f, x) = \frac{1}{A_n^\delta} \sum_{k=0}^{n} A_k^\delta a_k(f)p_k(x), \quad x \in [-1,1],
\]

\[
A_n^\delta = \frac{\Gamma(k+\delta+1)}{\Gamma(k+1)\Gamma(\delta+1)}, \quad N = 0, 1, \ldots
\]

For a given \( \beta > 0 \) we define,

\[
D(\beta)f \sim \sum_{k=1}^{\infty} k^\beta a_k(f)
\]

and we say \( D(\beta)f \in L_p(dw) \) if there exists \( \varphi \in L_p(dw) \) such that \( a_k(\varphi) = k^\beta a_k(f) \). It is known form \([17]\) that there is some \( \delta > 1/2 \) such that

\[
\|\sigma_n^\delta(f)\|_{L_p(dw)} \leq C_2 \|f\|_{L_p(dw)}, \quad 1 \leq p \leq \infty,
\]

with \( C_2 \) independent of \( N \).

For \( f \in L_p(dw) \) we define the generalized vallee Poussin means of the ortho-normal \( \{p_k(x)\} \) by

\[
\eta_k(f)(x) = \sum_{k=0}^{\infty} \eta_k \frac{k}{N} a_k(f)p_k(x), \quad \forall x \in [-1,1],
\]

where \( \eta(u) \in C^\alpha(R^+) \) is a nonnegative non-increasing function with \( \eta(u) = 1 \) for \( u \in [0,1] \) and \( \eta(u) = 0 \) for \( u > 2 \). Then \( \eta_k(f, x) = f(x), \quad f \in L_p(dw) \). Combing \([9]\) with \([18, Proposition 2.1]\), we have

\[
\|\eta_k(f)\|_{L_p(dw)} \leq C_2 \|f\|_{L_p(dw)}.
\]

For a given \( N > 1 \) we define the polynomial kernels by

\[
K_N(x, y) = \frac{1}{A_N} \sum_{k=0}^{\infty} p_k(x)p_k(y), \quad x, y \in [-1,1].
\]

For a given discrete set \( T \subset [-1,1] \), the corresponding RKHS \( \mathcal{H}_{K_N}^T \) is defined as (Ref. e.g.\([19]\)) the linear space of the set of functions \( \{K_N := K_N(x, t): t \in T\}, x \in X \) with the inner product given by

\[
\langle K_N(f), K_N(g) \rangle = \frac{1}{A_N} \sum_{k=0}^{\infty} \langle p_k(x)p_k(y), p_k(t)p_k(t) \rangle.
\]
We assume the distribution \( \rho(x,y) = \rho(y|x)w(x) \) with \( w(x) \) being Jacobi weight. Corresponds to the scheme (3) we have the following scheme,

\[
f_{i,j} := \arg \min_{f \in \mathcal{C}_{K_j}} \{ E(f) + \lambda \Vert f \Vert_{K_j, \pi} \},
\]

(12)

We define the projection operator \( \pi \) on the space of measurable function \( f : X \to R \) as

\[
\pi(f)(x) = \begin{cases} 
1, & \text{if } (x) > 1, \\
-1, & \text{if } (x) < -1, \\
(f(x)), & \text{if } -1 \leq f(x) \leq 1.
\end{cases}
\]

Since \( V(y \pi(f)(x)) \leq V(f(x)) \), we know that \( E(\pi(f)) \leq E(f) \). \( \pi(f) \) is measurable, then for any \( m \geq (4/\varepsilon^2 + 4 \log(pM \varepsilon^{-1})) \varepsilon \) and \( 0 < \delta < 1 \), with confidence at least \( 1 - 2\delta \), we have

\[
E(\pi(f_{i,j})) - E(f_{i,j}) \leq \tilde{c} \cdot m^{-\theta}, \quad \theta = \frac{p\beta}{2 + p\beta(2 - \alpha_p)}, \]

Theorem 1.2. Under the assumption of Theorem 2.1, we have for all \( 0 < \delta < 1/2 \), with confidence at least \( 1 - 2\delta \),

\[
\mathcal{R}(\pi(f_{i,j})) - \mathcal{R}(f_{i,j}) \leq \begin{cases} 
\tilde{c}m^{-\theta}, & \text{for } p = 1, \\
\tilde{c}m^{-\theta/2}, & \text{for } p > 1,
\end{cases}
\]

where \( 0 < \tilde{c} \leq 2c \) and \( \tilde{c} \) is the same as in Theorem 2.1.

II. THE APPROXIMATION ERROR

We estimate the approximation error for \( \eta_h(f_{i,j}) \), i.e.

\[
E(\eta_h(f_{i,j})) - E(f_{i,j}) \leq \eta_h(f_{i,j}) \Vert_{K_j, \pi}.
\]

Let \( \mathcal{P}_n \) be the set of all algebraic polynomials of order not exceeding \( n \). For \( p_n \in \mathcal{P}_n \), denoted by \( T_n = \{x_{n,i} \}_{i=1}^n \) the zeroes of \( p_n(x) \) arranged as

\[-1 < x_{n,1} < x_{n,2} \cdots < x_{n,n} < 1.\]

Then, holds the Gauss quadrature formula (Ref. e.g.\([15]\))

\[
\int_{-1}^1 p_n(x)w(x)dx = \sum_{i=1}^n \lambda_{n,i} p_n(x_{n,i}), \quad p_n \in \mathcal{P}_{2n-1}.
\]

(13)

For any \( p_n \in \mathcal{P}_n \), there holds the Nikolskii inequality (see. e.g.\([13],[16]\))

\[
\Vert p_n \Vert_{L_2(\tilde{\omega})} \leq C_n \Vert p_n \Vert_{L_2(\omega)}^{1-p \delta},
\]

(14)

where \( \alpha_p = \max(a,0), \ 1 \leq p \leq \infty \).

Proposition 2.1. Let \( \eta_h(f), \ K_j(x,y) \) be defined as (10) and (12) respectively. Then, for any \( f \in L_p(\omega) \) \( p \geq 1, \beta > 0 \),

\[
E(\eta_h(f_{i,j})) - E(f_{i,j}) \leq \eta_h(f_{i,j}) \Vert_{K_j, \pi} \leq M_{\alpha_p} \frac{\rho(\beta, \pi) \theta}{\rho(\beta, \pi)} + \lambda C_n \tilde{c}^{1-\delta},
\]

(15)

To prove Proposition 2.1, we firstly bound \( \eta_h(f_{i,j}) \Vert_{K_j, \pi} \).

Lemma 2.2. For any \( f \in L_p(\omega) \) there is \( C_r > 0 \) such that

\[
\Vert \eta_h(f) \Vert_{L_2(\omega)}^2 \leq C_r N^2 \tilde{c}^{1-\delta}, \quad \text{for } p = 1.
\]

(16)

Lemma 2.3 ([19]). Let \( D(\beta) \) be given (8). There is a constant \( M_0 > 0 \) such that

\[
\Vert \eta_h(f) - f \Vert_{L_2(\omega)} \leq M_0 \Vert D(\beta) \Vert_{L_2(\omega)}.
\]

(17)

Lemma 2.4 ([3], Theorem 25). If \( f : X \to R \) is measurable, then

\[
E(\pi(f)) - E(f_{i,j}) \leq \begin{cases} 
\rho(\beta, \pi) \theta / \rho(\beta, \pi), & \text{if } 1 \leq p \leq 2, \\
p(2-\beta)^{-1} \rho(\beta, \pi) \theta / \rho(\beta, \pi), & \text{if } p > 2.
\end{cases}
\]

(18)

Lemma 2.5. There are constants \( M_1 > 0, M_2 > 0, \ M_3 > 0 \) such that \( p \geq 1, \beta > 0 \),

\[
E(\eta_h(f_{i,j})) - E(f_{i,j}) \leq M_1 \frac{\rho(\beta, \pi) \theta}{\rho(\beta, \pi)},
\]

(19)

where \( M_1 = p(2-\beta)^{-1}(M_2^{-1} + 1)M_2, \ M_2 \leq 2N^{\beta} / M_1 \).

Now Proposition 2.1. can be derived from Lemma 2.2 and Lemma 2.5.

III. SAMPLE ERROR

Proposition 3.1. Let \( 0 < \varepsilon < 1/2 \), \( \lambda = \exp(-2m^\delta) \), \( \alpha_p = \min(p/2, 2/p) \) for \( p > 1 \) and \( \alpha [0,1] \) for \( p = 1 \), \( \beta > 0 \), \( N \leq m^{\delta} \). Then, for all \( m \geq (4/\varepsilon^2 + 4 \log(pM \varepsilon^{-1})) \varepsilon \) and \( 0 < \delta < 1 \), with confidence at least \( 1 - 2\delta \), there holds

\[
\begin{align*}
&\mathcal{E}(\pi(f_{i,j})) - \mathcal{E}(f_{i,j}) \leq \mathcal{E}(f_{i,j}) + \mathcal{E}(f_{i,j}) - \mathcal{E}(f_{i,j}) + c(m),
\end{align*}
\]

(20)

where
\[ c(m) = \frac{M_1}{2N^{m^2}} + \frac{2^{n+2} \log \frac{1}{\delta}}{3m} + \frac{2B_p \log \frac{1}{\delta}}{m^{1/2}} + 20 \left( \frac{2^{n+1} C}{3m^{1/2}} + \frac{2^{n+1} \log \frac{1}{\delta}}{3m} + \frac{8CB_p \log \frac{1}{\delta}}{m^{1/2}} \right)^{(1/2-n_1)}. \]

For simplicity, we divide the sample error in (13) for \( f = \eta_x(f_n) \) into two terms. The first term is \( A = [\mathcal{E}(\eta_x(f_n)) - \mathcal{E}(f_n)] - [\mathcal{E}(\eta_x(f_n)) - \mathcal{E}(f_n)] \), the second one is \( B = [\mathcal{E}(\pi(x,f_n)) - \mathcal{E}(f_n)] - [\mathcal{E}(\pi(x,f_n)) - \mathcal{E}(f_n)] \). For a compact set \( F \subset X \), we denote by \( \mathcal{N}(F,X) \) (see (23)) the set of functions in \( X \) that have more than a unique best approximation in \( F \) with respect to the norm, and \( \lambda'(f) = \sup \{ \lambda' \geq 1 : \lambda' f \in \mathcal{N}(F,X) \}. \)

**Lemma 3.2 (23).** Let \( 1 < p < \infty \) and set \( F \subset L_p(\mu) \) be a compact set of functions that are bounded by 1 put \( Y \) to be a random variable bounded by 1. If \( Y \in \mathcal{N}(F,L_p(\mu)) \), set \( \lambda'(Y) = \sup \{ \lambda' \geq 1 : \lambda' f + (1-\lambda')Y f \notin \mathcal{N}(F,X) \}. \) Then, for every \( f \in F \),

\[ E[V(Yf(x)) - V(f_n(x))]^2 \leq B_p \mathcal{E}(f) - \mathcal{E}(f_n)^{\gamma}, \]

(21)

where \( B_p = c(p) \inf_{1 \leq \lambda' \leq \lambda'} \lambda' - 1 \cdot c(p) \).

**Lemma 3.3.** For any \( 0 < \delta < 1 \), with confidence at least \( 1-\delta \), there holds

\[ A \leq 2^{n+3} \log \frac{1}{\delta} + \frac{2B_p \log \frac{1}{\delta}}{m^{1/2}} \frac{1}{3m} + \frac{2}{3} \left( \mathcal{E}(\eta_x(f_n)) - \mathcal{E}(f_n) \right). \]

(22)

**Lemma 3.4 ([1],[6]).** Let \( \xi \) be a random variable on probability space \( Z \) with mean \( E\xi = \mu \) and variance \( \sigma^2(\xi) = \sigma^2 \). If \( |\xi - \mu| \leq B \) almost everywhere, then for all \( \tau > 0 \),

\[ \text{Prob}_{x \in Z} \left\{ \frac{1}{m} \sum_{i=1}^m \xi(z_i) - \mu \geq \tau \right\} \leq \exp \left\{ - \frac{m\tau^2}{2(\sigma^2 + \frac{1}{3} B\tau)} \right\}. \]

(23)

To prove the result of sample error we need to extend (22) to a function set by means of the covering number. So let us recall some definitions. Let \( \mathcal{F} \) be a function set by means of the covering number. For simplicity, we set \( \mathcal{N}(\mathcal{F},r') \) is defined to be the minimal integer \( l \in \mathbb{N} \) such that there exist \( l \) balls with radius \( r' \) covering \( \mathcal{F} \). Denote the covering number of the unit ball \( B_r \) as \( \mathcal{N}(r') := \mathcal{N}(B_r,r') \), \( \forall r' > 0 \).

Take \( B_{r'} = \{ f \in \mathcal{H}_r : \| f \|_{K_x,\tau} \leq R \} \).

We recall some well-known results which will be used in our estimate for sample error.

**Lemma 3.5 ([1]).** Let \( E \) be a finite dimension Banach space with norm \( \| f \|_E \), \( r = \text{dim} E \), \( B_r = \{ f \in E : \| f \|_E \leq R \} \), then

\[ \log \mathcal{N}(B_r,r') \leq r \log \frac{4R}{r'} \]

The reproducing property implies \( \| f \|_u \leq \sup_{x \in X} K_y(x,y) \| f \|_{K_x,\tau} \leq 2N + 1 \), so Lemma 3.5 implies

\[ \log \mathcal{N}(B_{r'},r') \leq (2N + 1) \log \frac{4(2N + 1)R}{r'} \leq CN \log \frac{\sqrt{NR}}{r'} \]

(24)

The next known result we need is a quantitative version of the law of the large numbers

**Lemma 3.6 ([6]).** Let \( 0 \leq \alpha \leq 1 \), \( B > 0 \), \( c > 0 \), and \( G \) be a set of functions on \( Z \), such that for every \( g \in G \), \( E_g \geq 0 \), \( \text{Var}(g) \leq B \) and \( E_g^2 \leq c \text{Var}(g) \). Then for any \( \tau > 0 \),

\[ \text{Prob}_{x \in Z} \left\{ \sup_{g \in G} \frac{E_g}{m} \leq \tau \right\} \leq \mathcal{N}(G,\tau) \exp \left\{ - \frac{m\tau^2}{2(\alpha^2 + \frac{1}{3} B\tau)} \right\}. \]

(25)

By direct calculation it is not difficult to derive the following lemma.

**Lemma 3.7.** Let \( 0 < \zeta < \frac{1}{2} \). Then, for all \( m \geq 4(\zeta^2 + 4 \log (pM^{-\alpha}))^{1/2} \) and \( 0 < \delta < 1 \), with confidence at least \( 1-\delta \), we have

\[ \frac{1}{2} \| \mathcal{E}(\pi(x,f_n)) - \mathcal{E}(f_n) \| \leq \frac{1}{2} \| \mathcal{E}(\pi(x,f_n)) - \mathcal{E}(f'_n) \| + 20r \]

(26)

for all \( f \in B_{r'} \), where \( r' \) is equal to

\[ \frac{2^{n+1} C}{3m^{1/2}} + \frac{2^{n+1} \log(1/\delta)}{3m} + \frac{8CB'_p \log(1/\delta)}{m^{1/2}} \left( \frac{1}{3m} \right)^{(1/2-n_1)}. \]

(27)

**Proof of Proposition 3.1.** The result of Proposition 3.1 can be derived from Lemma 3.3 and Lemma 3.6.

### IV. ESTIMATE OF LEARNING RATES

Now we are in a position to present the rate of learning error. The desired learning rate can be derived by
combining the previous results on approximation error and sample error.

Proof of Theorem 1.1. Putting (17) and (22) into (16) with \( f = \eta_h(f_{i, j}) \), we have

\[
\mathcal{E}(\pi(f_{i, j})) - \mathcal{E}(f_{i, j}) \leq \lambda C_1^2 N^{2(1 - \frac{1}{p})} + \frac{M_1}{N^{\rho p}} + \frac{3M_1}{3m} + \frac{2}{3m} \left( \frac{2B_p \log \frac{1}{\delta}}{m^{\frac{1}{2} - \nu}} + \frac{2}{3m} \left( \frac{2B_p \log \frac{1}{\delta}}{m^{\frac{1}{2} - \nu}} \right) \right) + \frac{40}{3m} \left( \frac{2 \nu + 4B_p \log \frac{1}{\delta}}{m^{\frac{1}{2} - \nu}} \right),
\]

where \( c(m) \) given by (22), we have with confidence at least \( 1 - 2\delta \), that

\[
\mathcal{E}(\pi(f_{i, j})) - \mathcal{E}(f_{i, j}) \leq \lambda C_1^2 N^{2(1 - \frac{1}{p})} + \frac{M_1}{N^{\rho p}} + \frac{3M_1}{3m} + \frac{2}{3m} \left( \frac{2B_p \log \frac{1}{\delta}}{m^{\frac{1}{2} - \nu}} \right) + \frac{40}{3m} \left( \frac{2 \nu + 4B_p \log \frac{1}{\delta}}{m^{\frac{1}{2} - \nu}} \right).
\]

(28)

Since \( N \sim m^\gamma \) and \( \lambda = \exp\{-2m^\gamma\} \), we have

\[
\lambda N^{2(1 - \frac{1}{p})} \leq \frac{1}{\exp\{m^\gamma\}},
\]

then (28) can be rewritten as

\[
\mathcal{E}(\pi(f_{i, j})) - \mathcal{E}(f_{i, j}) \leq 2\lambda C_1^2 \left( \frac{1}{\exp\{m^\gamma\}} \right) + \frac{3M_1}{3m} + \frac{2}{3m} \left( \frac{2B_p \log \frac{1}{\delta}}{m^{\frac{1}{2} - \nu}} \right) + \frac{40}{3m} \left( \frac{2 \nu + 4B_p \log \frac{1}{\delta}}{m^{\frac{1}{2} - \nu}} \right).
\]

Take \( \xi = \frac{1}{2 + \rho \beta(2 - \alpha_p)} \), we have

\[
\mathcal{E}(\pi(f_{i, j})) - \mathcal{E}(f_{i, j}) \leq \tilde{c}m^{-\rho}.
\]

Proof of Theorem 1.2. Theorem 1.2 can be derived immediately by combining the result of Theorem 1.1 and the following relations between the expected risk and excess misclassification error:

\[
\mathcal{R}(\text{sgn}(f)) - \mathcal{R}(f) \leq \mathcal{E}(\pi(f)) - \mathcal{E}(f), \text{ for } p = 1,
\]

\[
\sqrt{2}(\mathcal{E}(\pi(f)) - \mathcal{E}(f)), \text{ for } p > 1,
\]

where \( f : X \rightarrow R \) is measurable, Ref.[3].

Acknowledgment

This work was supported in part by a grant from Natural Science Foundation of China (Grant No.10871226, 10971251) and Zhejiang Natural Science Foundation of Zhejiang Province(Grant No. Y6100096).

Ye Peixin is supported by the Program for New Century Excellent Talents in University of China.

References

**Baohuai Sheng** attended Baoji Normal College, Baoji, Shaanxi, from 1981 to 1985. He earned his BS degree in mathematical teaching from the department of mathematics in 1985. He earned his M.S. degree in basic mathematics from the department of mathematics of Hangzhou University in 1988. From 1988-2001 he worked towards the Doctor of Sciences in the applied mathematics at Xidian University, Xian, P.R.China, and earned his S. D. degree in March, 2001. From March, 2001, to September, 2003, he served as a Professor at the Ningbo University. Since September, 2003, he has been a faculty member of Shaoxing College of Arts and Sciences, Shaoxing, Zhejiang, P.R.China, where he serves currently as a Professor and the Chairman of Mathematical Department. His current research interests focus on the area of approximation theory, nonlinear optimization, learning theory.

**Peixin Ye** received the B.S. Degree and the M.S. degree in mathematics from Xiamen University, Fujian, China, in 1995 and 1998 respectively. He received the Ph.D. degree in mathematics from Beijing Normal University, in 2001. From 2001 to 2003, he worked in the Institute of mathematics, Chinese academy of sciences as a postdoctoral. He is currently a Full Professor at School of mathematical sciences, Nankai University. He has published more than 60 journal and conference papers. His current research interests include approximation theory, numerical analysis, quantum computing, machine learning and compressed sensing.