TAIWANESE JOURNAL OF MATHEMATICS
Vol. 18, No. 5, pp. 1633-1651, October 2014
DOI: 10.11650/tjm.18.2014.3929
This paper is available online at http://journal.taiwanmathsoc.org.tw

CLASSIFICATION WITH POLYNOMIAL KERNELS AND l^1-COEFFICIENT REGULARIZATION

Hongzhi Tong*, Di-Rong Chen and Fenghong Yang

Abstract. In this paper we investigate a class of learning algorithms for classification generated by regularization schemes with polynomial kernels and l^1-regularizer. The novelty of our analysis lies in the estimation of the hypothesis error. A Bernstein-Kantorovich polynomial is introduced as a regularizing function. Although the hypothesis spaces and the regularizers in the schemes are sample dependent, we prove the hypothesis error can be removed from the error decomposition with confidence. As a result, we derive some explicit learning rates for the produced classifiers under some assumptions.

1. INTRODUCTION

We consider binary classification algorithms generated by regularization schemes with general convex loss functions and polynomial kernels. Let X be a compact metric space (input space) and $Y = \{-1, 1\}$ (representing the two classes). Classification algorithms produce binary classifiers $\mathcal{C} : X \to Y$, which divide the input space into two classes. The misclassification error is used to measure the prediction power of a classifier $\mathcal{C}$. If ρ is a probability measure on $Z := X \times Y$, then the misclassification error for $\mathcal{C}$ is defined to be the probability of the event $\{\mathcal{C}(x) \neq y\}$:

$$\mathcal{R}(\mathcal{C}) := Prob\,\{\mathcal{C}(x) \neq y\} = \int_X \rho(y \neq \mathcal{C}(x)|x)d\rho_X.$$

Here ρ_X is the marginal distribution on X and $\rho(\cdot|x)$ is the conditional probability measure at $x \in X$ induced by ρ. It has been known from [6] the classifier which

Received October 19, 2013, accepted March 6, 2014.
Communicated by Chong Li.
2010 *Mathematics Subject Classification*: 68T05, 62J02.
Key words and phrases: Classification, Coefficient regularization, Polynomial kernels, Bernstein-Kantorovich polynomial, Learning rates.
This work was supported by National Natural Science Foundation of China under grants 11171014 and 11072274.
*Corresponding author.

minimizes the misclassification error is the Bayes rule. Recall the regression function of ρ :

$$f_\rho(x) = \int_Y y d\rho(y|x) = \rho(y=1|x) - \rho(y=-1|x), \quad x \in X.$$

Then the Bayes rule is given by the sign of the regression function $f_c := sgn(f_\rho)$. Here, for a function $f : X \to \mathbb{R}$, the sign function is defined as $sgn(f)(x) = 1$ if $f(x) \geq 0$ and $sgn(f)(x) = -1$ if $f(x) < 0$.

The classifiers considered here are induced by $f : X \to \mathbb{R}$ as $\mathcal{C} = sgn(f)$, where the real-valued functions f are generated from regularization schemes with general convex loss functions and polynomial kernels.

Definition 1.1. A continuous function V: $\mathbb{R} \to \mathbb{R}^+$ is called a classifying loss (function) if it is convex, differentiable at 0 with $V'(0) < 0$ and 1 is the smallest real for which the value of V is zero.

Examples of classifying loss include the hinge loss $V_h(t) = \max\{1-t, 0\}$ for the classical support vector machines (SVM) [15] classifier, and the least square loss $V_{ls}(t) = (1-t)^2$ (see [12]).

In this paper we consider the univariate input space $X = [0, 1]$. The polynomial kernel is defined by

$$K(x, u) := K_d(x, u) = (1 + xu)^d, \qquad \forall x, u \in X,$$

where d is the degree of kernel polynomial. We know from [4] that K is a Mercer kernel and the reproducing kernel Hilbert space $(\mathcal{H}_K, \|\cdot\|_K)$ associated with kernel K is the set of polynomials on X of degree at most d.

As ρ is unknown, the best classifier f_c can not be found directly. What we have in hand is a set of samples $\mathbf{z} := \{z_i\}_{i=1}^m = (x_i, y_i)_{i=1}^m \in Z^m$ independently drawn according to ρ. We call

$$\mathcal{E}_{\mathbf{z}}(f) := \mathcal{E}_{\mathbf{z}}^V(f) = \frac{1}{m}\sum_{i=1}^m V(y_i f(x_i))$$

the empirical error with respect to $\mathbf{z}$. Regularization learning schemes are implemented by minimizing a penalized version of the empirical error over a set of functions $\mathcal{H}$, called a hypothesis space, equipped with a penalty functional $\Omega : \mathcal{H} \to \mathbb{R}^+$, called a regularizer that reflects constraints imposed on functions from hypothesis space in various desirable forms.

With classifying loss V and polynomial kernel K, [22] analyzes a regularized classifier $sgn(\tilde{f}_{\mathbf{z},\lambda})$, where

$$\tilde{f}_{\mathbf{z},\lambda} := \arg\min_{f \in \mathcal{H}_K} \left\{\mathcal{E}_{\mathbf{z}}(f) + \lambda \|f\|_K^2\right\}. \tag{1.1}$$

Here $\lambda > 0$ is a regularization parameter. Particularly, when V is the hinge loss V_h, (1.1) becomes the classical SVM soft margin classifier (see [3]).

In this paper we shall consider a different regularization scheme. In our setting, the regularizer is rather than a reproducing kernel Hilbert space norm but a l^1-norm of the coefficients in the kernel ensembles. Let

$$\mathcal{H}_{K,\mathbf{z}} := \left\{ \sum_{i=1}^{m} a_i K_{x_i} : a_i \in \mathbb{R}, \ \ i = 1, 2, \cdots m \right\},$$

where $K_u(\cdot) = K(u, \cdot) = K(\cdot, u)$, and

$$\Omega_{\mathbf{z}}(f) := \inf \left\{ \sum_{i=1}^{m} |a_i| : f = \sum_{i=1}^{m} a_i K_{x_i} \right\}.$$

Then the regularized classifier with polynomial kernel K considered in this paper is given by $sgn(f_{\mathbf{z},\lambda})$, where $f_{\mathbf{z},\lambda}$ is a minimizer of the following optimization problem:

$$f_{\mathbf{z},\lambda} := \arg \min_{f \in \mathcal{H}_{K,\mathbf{z}}} \left\{ \mathcal{E}_{\mathbf{z}}(f) + \lambda \Omega_{\mathbf{z}}(f) \right\}. \tag{1.2}$$

Algorithms like (1.2) are also called coefficient regularization (see [11]). Recently, l^1-coefficient regularization has attracted much attention, the increasing interest is mainly brought by the progress of lasso in statistics [13] and compressive sensing in signal processing [2]. An essential difference between scheme (1.2) and (1.1) is the dependence of the hypothesis space and regularizer on samples $\mathbf{z}$. It raises a need of different methods for analyzing scheme (1.2). For instance, a key approach for scheme (1.1) used in [22] is an error decomposition which decomposes the total error into a sum of sample error and regularization error. However, this typical error decomposition technique does not apply for scheme (1.2) due to the dependence of $\mathcal{H}_{K,\mathbf{z}}$ and $\Omega_{\mathbf{z}}(\cdot)$ on $\mathbf{z}$. This was pointed out in [19] where a modified error decomposition was introduced by means of an extra hypothesis error. Under the framework established in [19], [20, 14] study least square regression and SVM regression with l^1-regularizer. The novelty of this paper is that we illustrate the hypothesis error can be removed with confidence by a special choice of the regularizing function for polynomial kernels. As a result, we derive some explicit learning rates for learning scheme (1.2) by estimating the sample error and regularization error respectively.

2. Error Decomposition

Define the generalization error associated with classifying loss V as

$$\mathcal{E}(f) := \mathcal{E}^V(f) = \int_Z V(yf(x)) d\rho.$$

Let f_ρ^V be a measurable function minimizing the generalization error, that is

$$f_\rho^V := \arg\min\{\mathcal{E}(f) : f \text{ is a measurable function on } X\}.$$

Since the smallest zero of V is 1, it is shown in [17] that f_ρ^V can be chosen such that $f_\rho^V(x) \in [-1, 1]$ for all $x \in X$.

Our goal is to estimate the excess misclassification error

$$\mathcal{R}(sgn(f_{\mathbf{z},\lambda})) - \mathcal{R}(f_c). \tag{2.1}$$

The following comparison theorem given by [3, 21] implies that the excess misclassification error can be bounded by the excess generalization error.

Proposition 2.1. *Let V be a classifying loss, then for any measurable function f,*

$$\mathcal{R}(sgn(f)) - \mathcal{R}(f_c) \le \begin{cases} \mathcal{E}(f) - \mathcal{E}(f_c) & if \ \ V(t) = (1-t)_+, \\ C_V\sqrt{\mathcal{E}(f) - \mathcal{E}(f_\rho^V)} & if \ \ V''(0) \ge 0, \end{cases}$$

where C_V is some constant dependent on V.

One always gets better estimates by making full use of the projection operator introduced in [1].

Definition 2.1. The projection operator π is defined on the space of measurable functions $f : X \to \mathbb{R}$ as

$$\pi(f)(x) := \begin{cases} 1, & if \ \ f(x) > 1, \\ -1, & if \ \ f(x) < -1, \\ f(x), & if -1 \le f(x) \le 1. \end{cases}$$

Trivially $sgn(\pi(f)) = sgn(f)$, Proposition 2.1 tells us

$$\mathcal{R}(sgn(f)) - \mathcal{R}(f_c) \le \begin{cases} \mathcal{E}(\pi(f)) - \mathcal{E}(f_c) & if \ \ V(t) = (1-t)_+, \\ C_V\sqrt{\mathcal{E}(\pi(f)) - \mathcal{E}(f_\rho^V)} & if \ \ V''(0) \ge 0. \end{cases} \tag{2.2}$$

Therefore, to estimate (2.1) it is sufficient for us to bound the excess generalization error $\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f_\rho^V)$. In addition, we can get immediately from Definition 1.1 that $V(y\pi(f)(x)) \le V(yf(x))$, so for any measurable function f,

$$\mathcal{E}(\pi(f)) \le \mathcal{E}(f), \quad \mathcal{E}_{\mathbf{z}}(\pi(f)) \le \mathcal{E}_{\mathbf{z}}(f). \tag{2.3}$$

Let ν be a Borel measure on X, we denote L_ν^p $(1 \le p < \infty)$ the measurable functions on X with norm $\|f\|_{L_\nu^p} := \left(\int_X |f(x)|^p d\nu\right)^{\frac{1}{p}} < \infty$. When ν is the Lebesgue measure, we simply denote L_ν^p as L^p. We also denote $C(X)$ as the space of continuous functions on X with the uniform norm $\|\cdot\|_\infty$.

Now we introduce the Bernstein-Kantorovich polynomials [9] that will play a key role in our analysis.

Definition 2.2. Let $f \in L^1$, the Bernstein-Kantorovich polynomial (of degree d) for f on X is given by

$$B_d(f,x) := \sum_{k=0}^{d} p_{d,k}(x)(d+1) \int_{k/d+1}^{(k+1)/d+1} f(u)du, \tag{2.4}$$

where $p_{d,k}(x) \equiv \binom{d}{k} x^k (1-x)^{d-k}, \quad k = 0, 1, \cdots, d$, are the Bernstein basis polynomials of degree d.

To formulate the error decomposition, we need to make use of a polynomial reproduction in the univariate case $X = [0,1]$.

Definition 2.3. Let T be a normed linear space with dual T^*. Given two subspaces $W \subseteq T$ and $U \subseteq T^*$, the set U is called a norming set of W if there exists some $c > 0$ so that

$$\sup_{u \in U, \|u\|=1} |u(w)| \geq c\|w\| \quad \forall w \in W.$$

With δ_x we denote the point evaluation functional at x, i.e.: $\delta_x(f) = f(x)$. The following proposition was given in [16], which was a reformulation of the result of [8].

Proposition 2.2. *If $\{x_1, x_2 \cdots, x_m\} \subset X$, W is a finite dimensional subspace of $C(X)$ and $U = span\{\delta_{x_i} : 1 \leq i \leq m\}$ is a norming set of W with norming constant $c \geq 1/2$. Then for every $w^* \in W^*$ with $\|w^*\| = 1$ there exist real numbers a_i with $\sum_{i=1}^m a_i w(x_i) = w^*(w)$ and $\sum_{i=1}^m |a_i| \leq 2$.*

Definition 2.4. A set $\{x_1, x_2 \cdots, x_m\} \subset X$ is said to be $\Delta-$dense if for any $x \in X$ there exists some $1 \leq i \leq m$ such that $|x - x_i| < \Delta$.

Lemma 2.1. *Let $\mathcal{P} = \mathcal{P}^d$ be the space of polynomials of degree d. If $\{x_1, x_2 \cdots, x_m\}$ is $\Delta-$dense in X with $\Delta \leq \frac{1}{4d^2}$, then $U = span\{\delta_{x_i} : 1 \leq i \leq m\}$ is a norming set of $\mathcal{P}$ with norming constant $c = 1/2$.*

Proof. For any $p \in \mathcal{P}$, there exists an $\bar{x} \in X$ with $p(\bar{x}) = \|p\|_\infty$. Since $\{x_1, x_2 \cdots, x_m\}$ is $\Delta-$dense in X, there is some $1 \leq i \leq m$ such that $|\bar{x} - x_i| < \Delta$. By Lagrange's Mean Value Theorem, there is at least a ζ, between $\bar{x}$ and x_i, such that

$$|p(\bar{x}) - p(x_i)| = |p'(\zeta)||\bar{x} - x_i|.$$

Applying the Markov inequality which is given for $p \in \mathcal{P}$ by

$$|p'(t)| \leq 2d^2\|p\|_\infty, \quad t \in [0,1],$$

we have with $\Delta \leq \frac{1}{4d^2}$,

$$|p(\bar{x})| - |p(x_i)| \leq 2d^2\|p\|_\infty \Delta \leq \frac{1}{2}\|p\|_\infty.$$

So

$$|\delta_{x_i}(p)| = |p(x_i)| \geq \frac{1}{2}\|p\|_\infty.$$

This proves the lemma. ∎

An immediate consequence of Lemma 2.1 and Proposition 2.2 is

Proposition 2.3. *If $\{x_1, x_2 \cdots, x_m\}$ is $\Delta-$dense in X with $\Delta \leq \frac{1}{4d^2}$, then there exist for every $x \in X$ real numbers $a_i(x)$ such that $\sum_{i=1}^m |a_i(x)| \leq 2$ and $\sum_{i=1}^m a_i(x)p(x_i) = p(x)$ for all $p \in \mathcal{P}$.*

Definition 2.5. The margin distribution ρ_X is said to satisfy condition L_τ with $1 \leq \tau < \infty$ if for some $c_\tau > 0$ and any interval $B(x, r) := \{u \in X : |u - x| < r\}$, one has

$$\rho_X(B(x, r)) \geq c_\tau r^\tau, \quad \forall x \in X, 0 < r \leq 1. \tag{2.5}$$

Proposition 2.4. *If ρ_X satisfies condition L_τ with $\tau \geq 1$, and $\{x_1, x_2 \cdots, x_m\}$ are samples independently drawn from ρ_X, for any $t > 1$, choosing*

$$\Delta = 2\left[1 + \left(\frac{1}{c_\tau}\right)^{\frac{1}{\tau}}\right]\left(\frac{\log m + t}{m}\right)^{\frac{1}{\tau}}.$$

Then $\{x_1, x_2 \cdots, x_m\}$ is $\Delta-$dense in X with confidence $1 - e^{-t}$.

Proof. Let N be the minimal $l \in \mathbb{N}$ such that there exist l open intervals with radius $\eta/2$ covering X, then $N \leq \frac{2}{\eta}$. If $\{B_j\}_{j=1}^N$ are the open intervals with radius $\eta/2$ covering X, by Definition 2.5, for each j the probability of the event $\{x_i\}_{i=1}^m \cap B_j = \varnothing$ is $(1 - \rho_X(B_j))^m \leq (1 - c_\tau(\frac{\eta}{2})^\tau)^m$. So the probability for $\{x_i\}_{i=1}^m \cap B_j = \varnothing$ to be true for at least one $j \in 1, \cdots, N$ is at most

$$N(1 - c_\tau(\frac{\eta}{2})^\tau)^m \leq \frac{2}{\eta}\exp\left\{-mc_\tau(\frac{\eta}{2})^\tau)\right\}.$$

This implies $\{x_i\}_{i=1}^m$ is $\eta-$dense in X with confidence at least $1-\frac{2}{\eta}\exp\left\{-mc_\tau\left(\frac{\eta}{2}\right)^\tau)\right\}$. So, if η satisfies

$$\log\left(\frac{\eta}{2}\right) + mc_\tau\left(\frac{\eta}{2}\right))^\tau \geq t, \tag{2.6}$$

then $\{x_i\}_{i=1}^m$ is $\eta-$dense in X with confidence $1 - e^{-t}$. What is left is to verify Δ satisfies (2.6). To this end, we consider the strictly increasing function h_1 on $(0, +\infty)$ defined by

$$h_1(\eta) = \log\left(\frac{\eta}{2}\right) + mc_\tau\left(\frac{\eta}{2}\right)^\tau.$$

Take $\tilde{\eta}$ to be the positive solution to the equation $h_1(\eta) = t$. If $\tilde{\eta} \geq 2\left(\frac{1}{m}\right)^{\frac{1}{\tau}}$, then

$$t = h_1(\tilde{\eta}) \geq -\frac{1}{\tau}\log m + mc_\tau \left(\frac{\tilde{\eta}}{2}\right)^\tau,$$

thus

$$\tilde{\eta} \leq 2\left(\frac{t + \frac{1}{\tau}\log m}{mc_\tau}\right)^{\frac{1}{\tau}} \leq \Delta.$$

If $\tilde{\eta} < 2\left(\frac{1}{m}\right)^{\frac{1}{\tau}}$, we can see $\tilde{\eta} \leq \Delta$ still holds. Therefore

$$h_1(\Delta) \geq h_1(\tilde{\eta}) = t.$$

It follows that Δ satisfies inequality (2.6) and the proof of Proposition 2.4 is completed. ■

Let

$$C_1 = \left[8\left(1 + \left(\frac{1}{c_\tau}\right)^{\frac{1}{\tau}}\right)\right]^\tau. \tag{2.7}$$

From Proposition 2.3 and 2.4, we can get

Corollary 2.1. *Suppose ρ_X satisfies condition L_τ with $\tau \geq 1$, and $\{x_i\}_{i=1}^m$ are samples independently drawn from ρ_X. For any $t > 1$, when*

$$m \geq C_1(\log m + t)d^{2\tau}, \tag{2.8}$$

then with confidence $1 - e^{-t}$, we can find numbers $a_i(x)$, for every $x \in X$ such that

$$\sum_{i=1}^m a_i(x)p(x_i) = p(x)$$

for all $p \in \mathcal{P}$ and

$$\sum_{i=1}^m |a_i(x)| \leq 2.$$

Theorem 2.1. *Suppose $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ is a set of samples independently drawn according to the measure ρ, and ρ_X satisfies condition L_τ with $\tau \geq 1$. If $|f(x)| \leq 1$ for all $x \in X$, then for any $t > 1$ and m satisfying (2.8), with confidence $1 - e^{-t}$, there holds*

$$B_d(f) \in \mathcal{H}_{K,\mathbf{z}}, \text{ and } \Omega_{\mathbf{z}}(B_d(f)) \leq 2 \cdot 18^d.$$

Proof. Let $u_j = \frac{j}{d}$, $j = 0, 1, \cdots, d$. Then

$$K_{u_j}(x) = (1+u_j x)^d = (1-x+(1+u_j)x)^d = \sum_{l=0}^d (1+u_j)^l p_{d,l}(x), \quad j = 0, 1, \cdots, d.$$

The Bernstein basis polynomial $p_{d,k}(x)$ with $k \in \{0, 1, \cdots, d\}$ can be written as

$$p_{d,k}(x) = \sum_{j=0}^{d} c_{k,j} K_{u_j}(x),$$

where $\{c_{k,j}\}_{j=0}^{d}$ is the solution to the linear system

$$\sum_{j=0}^{d} (1+u_j)^l c_{k,j} = \delta_{l,k}, \quad l = 0, 1, \cdots, d.$$

By Cramer's rule,

$$c_{k,j} = \frac{D_{k,j}}{D}, \quad j = 0, 1, \cdots, d$$

where D is the Vandermonde determinant

$$D := D(1+u_0, 1+u_1, \cdots, 1+u_d) = \begin{vmatrix} 1 & 1 & \cdots & 1 \\ 1+u_0 & 1+u_1 & \cdots & 1+u_d \\ \vdots & & \vdots & \vdots \\ (1+u_0)^d & (1+u_1)^d & \cdots & (1+u_d)^d \end{vmatrix}$$

and $D_{k,j}$ is the determinant obtained from D by replacing the jth column by e_{k+1}. Take

$$f_j(x) = D(1+u_0, \cdots, 1+u_{j-1}, x, 1+u_{j+1}, \cdots, 1+u_d).$$

We can see

$$\frac{f_j(x)}{D} = \frac{\sum_{k=0}^{d} D_{k,j} x^k}{D} = \sum_{k=0}^{d} c_{k,j} x^k = \prod_{l \neq j} \frac{x - (1+u_l)}{u_j - u_l}.$$

It follows that for $k = 0, 1, \cdots, d$,

$$\begin{aligned} c_{k,j} &= \frac{(-1)^{d-k}}{\prod_{l \neq j}(u_j - u_l)} \sum_{\substack{j_1 < j_2 \cdots < j_{d-k} \\ j_i \neq j}} (1+u_{j_1}) \cdots (1+u_{j_{d-k}}) \\ &= \frac{(-1)^{d-k}(-1)^{d-j} d^d}{j!(d-j)!} \sum_{\substack{j_1 < j_2 \cdots < j_{d-k} \\ j_i \neq j}} (1+\frac{j_1}{d}) \cdots (1+\frac{j_{d-k}}{d}). \end{aligned}$$

Hence

$$|c_{k,j}| \leq \frac{d^d}{d!} \binom{d}{j} \binom{d}{k} 2^{d-k}.$$

We thus can write

$$\begin{aligned} B_d(f,x) &= \sum_{k=0}^{d} p_{d,k}(x)(d+1)\int_{k/d+1}^{(k+1)/d+1} f(u)du \\ &= \sum_{k=0}^{d}(d+1)\int_{k/d+1}^{(k+1)/d+1} f(u)du \sum_{j=0}^{d} c_{k,j}K_{u_j}(x) \\ &= \sum_{j=0}^{d}\beta_j K_{u_j}(x), \end{aligned}$$

and

$$\begin{aligned} |\beta_j| &= \left|\sum_{k=0}^{d} c_{k,j}(d+1)\int_{k/d+1}^{(k+1)/d+1} f(u)du\right| \\ &\le \sum_{k=0}^{d}|c_{k,j}| \\ &\le \frac{d^d}{d!}\binom{d}{j}\sum_{k=0}^{d}\binom{d}{k}2^{d-k} \\ &= \frac{d^d}{d!}\binom{d}{j}3^d. \end{aligned}$$

Since $K_x(\cdot)\in\mathcal{P}$ for any fixed $x\in X$, by Corollary 2.1, with confidence $1-e^{-t}$, we have for $j=0,1,\cdots,d$,

$$K_{u_j}(x)=K_x(u_j)=\sum_{i=1}^{m}a_i(u_j)K_x(x_i)=\sum_{i=1}^{m}a_i(u_j)K_{x_i}(x),$$

and

$$\sum_{i=1}^{m}|a_i(u_j)|\le 2.$$

Therefore,

$$B_d(f,x)=\sum_{j=0}^{d}\beta_j K_{u_j}(x)=\sum_{i=1}^{m}\sum_{j=0}^{d}a_i(u_j)\beta_j K_{x_i}(x)\in\mathcal{H}_{K,\mathbf{z}},$$

and

$$\begin{aligned} \Omega_{\mathbf{z}}(B_d(f)) &\le \sum_{i=1}^{m}|\sum_{j=0}^{d}a_i(u_j)\beta_j| \le 2\sum_{j=0}^{d}|\beta_j| \\ &\le 2\sum_{j=0}^{d}\frac{d^d}{d!}\binom{d}{j}3^d = 2\cdot 6^d\frac{d^d}{d!}\le 2\cdot 18^d. \end{aligned}$$

The last inequality follows from the Stirling formula. ■

Taking $B_d(f_\rho^V)$ as the regularizing function, we can now present the error decomposition as following.

Theorem 2.2. *If $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ are independently drawn according to ρ, and ρ_X satisfies condition L_τ with $\tau \geq 1$. Then for any $t > 1$ and m satisfying (2.8), with confidence $1 - e^{-t}$,*

$$\begin{aligned}\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f_\rho^V) \leq\ & [\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda}))\\ & +\mathcal{E}_{\mathbf{z}}(B_d(f_\rho^V)) - \mathcal{E}(B_d(f_\rho^V))]\\ & +[\mathcal{E}(B_d(f_\rho^V)) - \mathcal{E}(f_\rho^V) + \lambda\Omega_{\mathbf{z}}(B_d(f_\rho^V))].\end{aligned} \tag{2.9}$$

Proof. Since $f_\rho^V(x) \in [-1, 1]$ for all $x \in X$, according to Theorem 2.1, we know that $B_d(f_\rho^V) \in \mathcal{H}_{K,\mathbf{z}}$ with confidence $1 - e^{-t}$. So under the same confidence,

$$\begin{aligned}& \mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f_\rho^V)\\ \leq\ & \mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f_\rho^V) + \lambda\Omega_{\mathbf{z}}(f_{\mathbf{z},\lambda})\\ =\ & [\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda}))] + [(\mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda})) + \lambda\Omega_{\mathbf{z}}(f_{\mathbf{z},\lambda}))\\ & -(\mathcal{E}_{\mathbf{z}}(B_d(f_\rho^V)) + \lambda\Omega_{\mathbf{z}}(B_d(f_\rho^V)))]\\ & +[\mathcal{E}_{\mathbf{z}}(B_d(f_\rho^V)) - \mathcal{E}(B_d(f_\rho^V))]\\ & +[\mathcal{E}(B_d(f_\rho^V)) - \mathcal{E}(f_\rho^V) + \lambda\Omega_{\mathbf{z}}(B_d(f_\rho^V))].\end{aligned} \tag{2.10}$$

It follows from (2.3) that $\mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda})) \leq \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda})$. This in connection with the definition of $f_{\mathbf{z},\lambda}$ implies the second term of (2.10) is at most zero. So the theorem is proved. ■

The second term in (2.10) with a regularizing function $f_\lambda \in \mathcal{H}_K$ is called in [19] hypothesis error, caused by the sample dependence of the hypothesis space $\mathcal{H}_{K,\mathbf{z}}$ which need not contain the regularizing function f_λ. Dealing with hypothesis error is usually the key and difficult step in the analysis of algorithms established in a data dependent hypothesis space (see [14, 20]). However, Theorem 2.2 ensures us the hypothesis error can be discarded with high confidence, which is mainly attributed to the special structure of polynomial kernels.

3. Error Analysis

As usual, the first and second term on the right side of (2.9) are respectively called sample error and regularization error. In this section, we shall provide some bounds for them separately.

3.1. Estimating $\mathcal{E}(B_d(f_\rho^V)) - \mathcal{E}(f_\rho^V)$

Since V is convex, its left, right derivatives $V'_-,\ V'_+$ exist.

Proposition 3.1. *Let V be a classifying loss function and $M_0 := \max\{|V'_\pm(-1)|, |V'_\pm(1)|\}$, then there holds*

$$\mathcal{E}(B_d(f_\rho^V)) - \mathcal{E}(f_\rho^V) \le M_0 \|B_d(f_\rho^V) - f_\rho^V\|_{L^1_{\rho_X}}. \tag{3.1}$$

Proof. Since $f_\rho^V(x) \in [-1,1]$, we can see that $|B_d(f_\rho^V, x)| \le 1$ for each $x \in X$. By Theorem 4 in [17],

$$\mathcal{E}(B_d(f_\rho^V)) - \mathcal{E}(f_\rho^V) \le \|V'_\pm\|_{L^\infty[-1,1]} \|B_d(f_\rho^V) - f_\rho^V\|_{L^1_{\rho_X}}.$$

The convexity of V implies that the one-side derivatives $V'_+,\ V'_-$ are both nondecreasing, this proves the proposition. ∎

In order to estimate $\|B_d(f_\rho^V) - f_\rho^V\|_{L^1_{\rho_X}}$, we give the following definition which was discussed in [4].

Definition 3.1. We call $\mathcal{D}_{\rho_X}$ the distortion of ρ_X (with respect to the Lebesgue measure), if $\mathcal{D}_{\rho_X}$ is the operator norm $\|J\|$ where J is the identity mapping

$$L^1 \xrightarrow{J} L^1_{\rho_X}.$$

$\mathcal{D}_{\rho_X}$ measures how much ρ_X distorts the Lebesgue measure. It is often reasonable to suppose that the distortion $\mathcal{D}_{\rho_X}$ is finite. Therefore

$$\|B_d(f_\rho^V) - f_\rho^V\|_{L^1_{\rho_X}} \le \mathcal{D}_{\rho_X} \|B_d(f_\rho^V) - f_\rho^V\|_{L^1}. \tag{3.2}$$

It has been known from the knowledge of approximation theory that approximation by Bernstein-Kantorovich polynomials can be characterized by the modulus of smoothness of the functions they approximate.

Definition 3.2. Let $\varphi(x) = \sqrt{x(1-x)}$ and

$$\Delta^2_{h\varphi} f(x) = \begin{cases} f(x - h\varphi(x)) - 2f(x) + f(x + h\varphi(x)), & if\ \ x \pm h\varphi(x) \in [0,1], \\ 0, & \text{otherwise.} \end{cases}$$

Then the modulus of smoothness of $f \in L^1$ is defined as

$$\omega^2_\varphi(f, r) := \sup_{0<h\le r} \|\Delta^2_{h\varphi} f\|_{L^1}.$$

From Theorem 9.3.2 in [7] we know that for any $f \in L^1$,

$$\|B_d(f) - f\|_{L^1} \le C_2 \left[\omega_\varphi^2(f, \frac{1}{\sqrt{d}}) + \frac{\|f\|_{L^1}}{d}\right],$$

where C_2 is a constant independent of f and d. This together with (3.1) and (3.2) implies

Proposition 3.2. *If $\mathcal{D}_{\rho_X} < \infty$, $\omega_\varphi^2(f_\rho^V, r) = O(r^{2s})$, $(0 < s \le 1)$. Then there exists a constant C_3 independent of d, such that*

$$\mathcal{E}(B_d(f_\rho^V)) - \mathcal{E}(f_\rho^V) \le C_3 \mathcal{D}_{\rho_X} d^{-s}.$$

3.2. Estimating $[\mathcal{E}_{\mathbf{z}}(B_d(f_\rho^V)) - \mathcal{E}_{\mathbf{z}}(f_\rho^V)] - [\mathcal{E}(B_d(f_\rho^V)) - \mathcal{E}(f_\rho^V)]$

For a measurable function $f : Z \to \mathbb{R}$, denote $\mathrm{E}f := \int_Z f(z)d\rho$. The following definition is a variance-expectation condition for the pair (V, ρ), which has been generally used to achieve sharp estimation of the sample error.

Definition 3.3. A variance power α of the pair (V, ρ) is a number in $[0, 1]$ such that for any $f : X \to [-1, 1]$, there exists some constant $c_\alpha > 0$ satisfying

$$\mathrm{E}[V(yf(x)) - V(yf_\rho^V(x))]^2 \le c_\alpha [\mathcal{E}(f) - \mathcal{E}(f_\rho^V)]^\alpha. \tag{3.3}$$

Remark 3.1. For $V = V_{ls}$, the power can be taken as $\alpha = 1$ (see [5]). For the hinge loss $V = V_h$, one can take $\alpha = \frac{q}{q+1}$, when a Tsybakov noise condition with exponent $q \ge 0$ is satisfied (see [10]). In general, (3.3) always holds for $\alpha = 0$ and $c_\alpha = (V(-1))^2$.

To complete the estimation, we need to use the following one-side Bernstein inequality (see [4]).

Let ξ be a random variable on a probability space Z with mean $\mathbb{E}\xi = \mu$ and variance $\sigma^2(\xi) = \sigma^2$. If $|\xi - \mu| \le B$ almost everywhere, then for all $\eta > 0$,

$$Prob_{\mathbf{z} \in Z^m} \left\{ \frac{1}{m} \sum_{i=1}^m \xi(z_i) - \mu \ge \eta \right\} \le \exp\left\{ -\frac{m\eta^2}{2(\sigma^2 + \frac{1}{3}B\eta)} \right\}.$$

Proposition 3.3. *If (3.3) holds, then for any $t > 1$, with the confidence $1 - e^{-t}$,*

$$\begin{aligned} &[\mathcal{E}_{\mathbf{z}}(B_d(f_\rho^V)) - \mathcal{E}_{\mathbf{z}}(f_\rho^V)] - [\mathcal{E}(B_d(f_\rho^V)) - \mathcal{E}(f_\rho^V)] \\ &\le \frac{4V(-1)t}{3m} + \left(\frac{2c_\alpha t)}{m}\right)^{1/(2-\alpha)} + \mathcal{E}(B_d(f_\rho^V)) - \mathcal{E}(f_\rho^V). \end{aligned}$$

Proof. Let $\xi = V(yB_d(f_\rho^V, x)) - V(yf_\rho^V(x))$. Since $|f_\rho^V(x)| \le 1$ and $|B_d(f_\rho^V, x)| \le 1$, we know by the monotonicity of V that $|\xi| \le V(-1)$. Hence $\mathrm{E}\xi = \mathcal{E}(B_d(f_\rho^V)) - \mathcal{E}(f_\rho^V)$, $|\xi - \mathrm{E}\xi| \le 2V(-1)$, and (3.3) yields $\sigma^2(\xi) \le \mathrm{E}(\xi^2) \le c_\alpha(\mathrm{E}\xi)^\alpha$. Applying the one-side Bernstein inequality to ξ, we find that for every $\eta > 0$,

$$Prob_{\mathbf{z}\in Z^m}\left\{\frac{1}{m}\sum_{i=1}^m \xi(z_i) - \mathrm{E}\xi \ge \eta\right\} \le \exp\left\{-\frac{m\eta^2}{2(c_\alpha(\mathrm{E}\xi)^\alpha + \frac{2}{3}V(-1)\eta)}\right\}.$$

So for any $t > 1$, with confidence $1 - e^{-t}$,

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^m \xi(z_i) - \mathrm{E}\xi &\le \frac{4V(-1)t}{3m} + \sqrt{\frac{2tc_\alpha(\mathrm{E}\xi)^\alpha}{m}} \\ &\le \frac{4V(-1)t}{3m} + \frac{\alpha}{2}\mathrm{E}\xi + (1-\frac{\alpha}{2})\left(\frac{2tc_\alpha}{m}\right)^{\frac{1}{2-\alpha}} \\ &\le \frac{4V(-1)t}{3m} + \left(\frac{2tc_\alpha}{m}\right)^{\frac{1}{2-\alpha}} + \mathrm{E}\xi.\end{aligned}$$

Here the second inequality follows from an elementary inequality

$$ab \le \frac{1}{p}a^p + \frac{1}{q}b^q, \quad \forall a, b > 0,\ p, q > 1,\ \frac{1}{p} + \frac{1}{q} = 1. \tag{3.4}$$

This proves the proposition. ∎

3.3. Estimating $[\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f_\rho^V)] - [\mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}_{\mathbf{z}}(f_\rho^V)]$

The function $f_{\mathbf{z},\lambda}$ changed with the sample $\mathbf{z}$ runs over a set of functions in $\mathcal{H}_K$, so we need a uniform probability inequality which involves the complexity of $\mathcal{H}_K$ described by means of the covering number.

Definition 3.4. For a subset $\mathcal{F}$ of a metric space and $\eta > 0$, the covering number $\mathcal{N}(\mathcal{F}, \eta)$ is defined to be the minimal integer $l \in \mathbb{N}$ such that there exist l balls with radius η covering $\mathcal{F}$.

We can see that the dimension of $\mathcal{H}_K$ is $d+1$. For any $f \in \mathcal{H}_K$, denote

$$\|f\| := \inf\left\{\sum_{j=1}^{d+1}|a_j|, f = \sum_{j=1}^{d+1} a_j K_{u_j}, u_j \in X\right\}$$

and $\mathcal{B}_R := \{f \in \mathcal{H}_K, \|f\| \le R\}$. It is easy to see that $\mathcal{B}_R$ is a subset of $C(X)$, and

$$\|f\|_\infty \le \sup_{x,u\in X}|K(x,u)| \cdot \|f\| \le 2^d R.$$

By Proposition 5 in Chapter I of [4], we have

$$\log \mathcal{N}(\mathcal{B}_R, \eta) \le (d+1)\log\left(\frac{4\cdot 2^d R}{\eta}\right). \tag{3.5}$$

The following lemma is adopted from [18], it can been seen as a uniform law of large numbers for a class of functions.

Lemma 3.1. *Let $0 \le \alpha \le 1, B > 0, c \ge 0$, and $\mathcal{G}$ be a set of functions on Z such that for every $g \in \mathcal{G}$, $\mathrm{E}g \ge 0$, $|g - \mathrm{E}g| \le B$ almost everywhere and $\mathrm{E}(g^2) \le c(\mathrm{E}g)^\alpha$. Then for every $\eta > 0$,*

$$Prob_{\mathbf{z}\in Z^m}\left\{\sup_{g\in\mathcal{G}} \frac{\mathrm{E}g - \frac{1}{m}\sum_{i=1}^m g(z_i)}{\sqrt{(\mathrm{E}g)^\alpha + \eta^\alpha}} > 4\eta^{1-\frac{\alpha}{2}}\right\} \le \mathcal{N}(\mathcal{G},\eta)\exp\left\{-\frac{m\eta^{2-\alpha}}{2(c+\frac{1}{3}B\eta^{1-\alpha})}\right\}.$$

Applying Lemma 3.1 to the following function set:

$$\mathcal{F}_R := \{V(y\pi(f)(x)) - V(yf_\rho^V(x)) : f \in \mathcal{B}_R\},$$

we can find

Proposition 3.4. *Let $R > 0$, if (3.3) holds, then for every $\eta > 0$,*

$$Prob_{\mathbf{z}\in Z^m}\left\{\sup_{f\in\mathcal{B}_R} \frac{[\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f_\rho^V)] - [\mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}_{\mathbf{z}}(f_\rho^V)]}{\sqrt{[\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f_\rho^V)]^\alpha + \eta^\alpha}} \le 4\eta^{1-\frac{\alpha}{2}}\right\}$$
$$\ge 1 - \exp\left\{(d+1)\log\left(\frac{4\cdot 2^d M_0 R}{\eta}\right) - \frac{m\eta^{2-\alpha}}{2(c_\alpha + \frac{2}{3}V(-1)\eta^{1-\alpha})}\right\}.$$

Here M_0 is given in Proposition 3.1.

Proof. Each function $g \in \mathcal{F}_R$ has the form $g(z) = g(x,y) = V(y\pi(f)(x)) - V(yf_\rho^V(x))$ with some $f \in \mathcal{B}_R$. Hence $\mathrm{E}g = \mathcal{E}(\pi(f)) - \mathcal{E}(f_\rho^V) \ge 0$, $\frac{1}{m}\sum_{i=1}^m g(z_i) = \mathcal{E}_{\mathbf{z}}(\pi(f)) - \mathcal{E}_{\mathbf{z}}(f_\rho^V)$. Furthermore,

$$\|g\|_\infty \le V(-1), \quad |g - \mathrm{E}g| \le 2V(-1).$$

(3.3) tells us $\mathrm{E}(g^2) \le C_\alpha(\mathrm{E}g)^\alpha$. Now applying Lemma 3.1 to $\mathcal{F}_R$, we have

$$Prob_{\mathbf{z}\in Z^m}\left\{\sup_{f\in\mathcal{B}_R} \frac{[\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f_\rho^V)] - [\mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}_{\mathbf{z}}(f_\rho^V)]}{\sqrt{[\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f_\rho^V)]^\alpha + \eta^\alpha}} > 4\eta^{1-\frac{\alpha}{2}}\right\}$$
$$\le \mathcal{N}(\mathcal{F}_R,\eta)\exp\left\{-\frac{m\eta^{2-\alpha}}{2(c_\alpha + \frac{2}{3}V(-1)\eta^{1-\alpha})}\right\}.$$

What is left is to bound the covering number. Observe that for any $f_1, f_2 \in \mathcal{B}_R$ and $(x, y) \in Z$,

$$\begin{aligned}
&|[V(y\pi(f_1)(x)) - V(yf_\rho^V(x))] - [V(y\pi(f_2)(x)) - V(yf_\rho^V(x))]| \\
&= |V(y\pi(f_1)(x)) - V(y\pi(f_2)(x))| \\
&\le |V'_+(-1)|\, |\pi(f_1)(x) - \pi(f_2)(x)| \\
&\le M_0\|f_1 - f_2\|_\infty.
\end{aligned}$$

This in connection with (3.5) means that

$$\log \mathcal{N}(\mathcal{F}_R, \eta) \le \log \mathcal{N}(\mathcal{B}_R, \frac{\eta}{M_0}) \le (d+1)\log(\frac{4 \cdot 2^d M_0 R}{\eta}).$$

So the proposition is proved. ■

We now need to find a ball $\mathcal{B}_R$ containing $f_{\mathbf{z},\lambda}$.

Lemma 3.2. *For all $\lambda > 0$ and $\mathbf{z} \in Z^m$, one has*

$$\|f_{\mathbf{z},\lambda}\| \le \frac{V(0)}{\lambda}. \tag{3.6}$$

Proof. By taking $f = 0$ in (1.2), one can see that

$$\lambda\|f_{\mathbf{z},\lambda}\| \le \lambda\Omega_{\mathbf{z}}(f_{\mathbf{z},\lambda}) \le \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda}) + \lambda\Omega_{\mathbf{z}}(f_{\mathbf{z},\lambda}) \le \mathcal{E}_{\mathbf{z}}(0) = V(0).$$ ■

Proposition 3.5. *Let $0 < \theta < 1/2, \lambda = e^{-4d}, d = (Cm)^\theta$ with some constant $C > 0$ independent of m. If (3.3) holds, then for all $t > 1$ and $m \ge \{1/\theta^2 + \log(4M_0V(0))\}^{1/\theta}$, with confidence $1 - e^{-t}$, there holds*

$$\begin{aligned}
&[\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f_\rho^V)] - [\mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}_{\mathbf{z}}(f_\rho^V)] \\
&\le \frac{1}{2}[\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f_\rho^V)] + C_4 t m^{-\frac{1-2\theta}{2-\alpha}},
\end{aligned}$$

where C_4 is a constant independent of m or t.

Proof. Taking $R = \frac{V(0)}{\lambda}$. By (3.6), $f_{\mathbf{z},\lambda} \in \mathcal{B}_R$ for all $\mathbf{z} \in Z^m$. Choose η^* to be the positive solution to the following equation

$$h_2(\eta) := \frac{m\eta^{2-\alpha}}{2(c_\alpha + \frac{2}{3}V(-1)\eta^{1-\alpha})} - (d+1)\log\left(\frac{4M_0V(0)2^d}{\lambda\eta}\right) = t. \tag{3.7}$$

Then Proposition 3.4 implies that with confidence $1 - e^{-t}$,

$$
\begin{aligned}
&[\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f_\rho^V)] - [\mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}_{\mathbf{z}}(f_\rho^V)] \\
\le\ & 4\eta^{*1-\alpha/2}\sqrt{[\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f_\rho^V)]^\alpha + \eta^{*\alpha}} \\
\le\ & 4\eta^* + 4\eta^{*1-\alpha/2}[\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f_\rho^V)]^{\alpha/2} \\
\le\ & 4\eta^* + \frac{\alpha}{2}[\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f_\rho^V)] + (1-\frac{\alpha}{2})4^{2/(2-\alpha)}\eta^* \\
\le\ & 20\eta^* + \frac{1}{2}[\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f_\rho^V)].
\end{aligned}
$$

Here in the third inequality we have used the elementary inequality (3.4) again.

It remains to estimate η^*. Let $\beta := \frac{1-2\theta}{2-\alpha}$. If $\eta^* \ge m^{-\beta}$, putting $\lambda = e^{-4d}, d = (Cm)^\theta$ into (3.7), we can see that

$$
\begin{aligned}
h_2(\eta^*) \ge\ & \frac{m\eta^{*2-\alpha}}{2(c_\alpha + \frac{2}{3}V(-1)\eta^{*1-\alpha})} - ((Cm)^\theta + 1)\left[\log(4M_0V(0))\right. \\
& \left. +(Cm)^\theta \log 2 + 4(Cm)^\theta + \beta\log m\right].
\end{aligned}
$$

Since $m \ge \{1/\theta^2 + \log(4M_0V(0))\}^{1/\theta}$, we have $\log(4M_0V(0)) \le m^\theta$ and

$$
\beta\log m < \log m < m^\theta. \tag{3.8}
$$

In fact, let $F(x) = x^\theta - \log x$, then $F'(x) = \frac{1}{x}(\theta x^\theta - 1)$. It means $F(x)$ is an increasing function when $x^\theta > \frac{1}{\theta^2}$, and $F\left((1/\theta^2)^{1/\theta}\right) = \frac{1}{\theta}\left(\frac{1}{\theta} + 2\log\theta\right) > 0$. So (3.8) holds.

Therefore,

$$
t = h_2(\eta^*) \ge \frac{m\eta^{*2-\alpha}}{2(c_\alpha + \frac{2}{3}V(-1)\eta^{*1-\alpha})} - (1+C^\theta)(2+(4+\log 2)C^\theta)m^{2\theta}. \tag{3.9}
$$

Denote $\tilde{C} := (1+C^\theta)(2+(4+\log 2)C^\theta)$, then (3.9) can be rewritten as

$$
\eta^{*2-\alpha} - \frac{4}{3}V(-1)\left[\frac{t}{m} + \tilde{C}m^{2\theta-1}\right]\eta^{*1-\alpha} - 2c_\alpha\left[\frac{t}{m} + \tilde{C}m^{2\theta-1}\right] \le 0.
$$

It implies that

$$
\begin{aligned}
\eta^* &\le \max\left\{\frac{8}{3}V(-1)\left(\frac{t}{m} + \tilde{C}m^{2\theta-1}\right), \left(4c_\alpha\left(\frac{t}{m} + \tilde{C}m^{2\theta-1}\right)\right)^{\frac{1}{2-\alpha}}\right\} \\
&\le \tilde{C}_4 t m^{-\frac{1-2\theta}{2-\alpha}},
\end{aligned} \tag{3.10}
$$

where $\tilde{C}_4 := 1 + \frac{8}{3}V(-1)(1+\tilde{C}) + \left(4c_\alpha(1+\tilde{C})\right)^{\frac{1}{2-\alpha}}$.

If $\eta^* < m^{-\beta}$, (3.10) still holds. So the proposition follows by taking $C_4 = 20\tilde{C}_4$. ∎

4. Learning Rate

Combining the estimations in the last section, we can derive some explicit learning rates for scheme (1.2) by choosing suitable values of λ and d.

Theorem 4.1. *Suppose $\mathcal{D}_{\rho_X} < \infty$, $\omega_\varphi^2(f_\rho^V, r) = O(r^{2s})$ for some $0 < s \leq 1$, ρ_X satisfies condition L_τ with $\tau \geq 1$ and (3.3) holds. Let $\theta = \frac{1}{1+2\tau}$, $d = \left(\frac{m}{(2C_1)^{1/(1-\theta)}}\right)^\theta$, $\lambda = e^{-4d}$. Then for any $t > 1$, when $m \geq \{t+1/\theta^2+\log(M_0V(0))\}^{1/\theta}$, with confidence $1-3e^{-t}$, we have*

$$\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f_\rho^V) \leq C_5 t m^{-\min\left\{\frac{2\tau-1}{(2-\alpha)(1+2\tau)}, \frac{s}{1+2\tau}\right\}},$$

where $C_5 = 2(C_4 + \frac{4V(-1)}{3} + (2c_\alpha)^{1/(2-\alpha)}) + 4(C_3\mathcal{D}_{\rho_X}+1)(2C_1)^{\frac{\theta s}{1-\theta}}$.

Proof. Since $m \geq \{t + 1/\theta^2 + \log(M_0V(0))\}^{1/\theta}$, we know from (3.8)

$$m^\theta > t, \quad m^\theta > \log m.$$

Hence

$$C_1(\log m + t)d^{2\tau} \leq 2C_1 m^\theta \left(\frac{m}{(2C_1)^{1/(1-\theta)}}\right)^{\theta\cdot\frac{1-\theta}{\theta}} = m.$$

Now by Theorem 2.2, with confidence $1-e^{-t}$, (2.9) holds. Putting Theorem 2.1, Proposition 3.2, 3.3 and Proposition 3.5 with $C = (2C_1)^{1/(\theta-1)}$ into the right side of (2.9), we find that with confidence $1-3e^{-t}$,

$$\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f_\rho^V) \leq \frac{1}{2}[\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f_\rho^V)] + C_4 t m^{-\frac{1-2\theta}{2-\alpha}}$$

$$+\frac{4V(-1)t}{3m} + \left(\frac{2c_\alpha t)}{m}\right)^{1/(2-\alpha)} + 2C_3\mathcal{D}_{\rho_X}d^{-s} + 2\left(\frac{18}{e^4}\right)^d.$$

Note that $\left(\frac{18}{e^4}\right)^d \leq \left(\frac{1}{2}\right)^d \leq d^{-s} = (2C_1)^{\frac{\theta s}{1-\theta}} m^{-s\theta}$, we have with same confidence

$$\begin{aligned}\mathcal{E}(\pi(f_{\mathbf{z},\lambda})) - \mathcal{E}(f_\rho^V) &\leq 2(C_4 + \tfrac{4V(-1)}{3} + (2c_\alpha)^{1/(2-\alpha)})tm^{-\frac{1-2\theta}{2-\alpha}} \\ &\quad +4(C_3\mathcal{D}_{\rho_X}+1)(2C_1)^{\frac{\theta s}{1-\theta}}m^{-\theta s} \\ &\leq C_5 t m^{-\min\left\{\frac{1-2\theta}{2-\alpha}, \theta s\right\}} \\ &= C_5 t m^{-\min\left\{\frac{2\tau-1}{(2-\alpha)(1+2\tau)}, \frac{s}{1+2\tau}\right\}}.\end{aligned}$$

■

Theorem 4.1 in connection with the relation (2.2) yields the learning rates with respect to misclassification error.

Corollary 4.1. *For all $0 < \delta < 1$, if the conditions in Theorem 4.1 are satisfied with $t = \log(3/\delta)$, then with confidence $1 - \delta$, there holds*

$$\mathcal{R}(sgn(f_{\mathbf{z},\lambda})) - \mathcal{R}(f_c) \leq \begin{cases} \mathcal{C}_5 \log(3/\delta) m^{-\gamma} & if \;\; V(t) = (1-t)_+, \\ \mathcal{C}_V \sqrt{C_5} \sqrt{\log(3/\delta)}\, m^{-\gamma/2} & if \;\; V''(0) \geq 0. \end{cases}$$

where $\gamma = \min\left\{ \frac{2\tau - 1}{(2-\alpha)(1+2\tau)}, \frac{s}{1+2\tau} \right\}$.

References

1. P. L. Bartlett, The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network, *IEEE Trans. Inform. Theory*, **44** (1998), 525-536.
2. E. J. Candès, J. Romberg and T. Tao, Robust uncertainty principles: exact signal reconstraction for highly incomplete frequency information, *IEEE Trans. Inform. Theory*, **52** (2006), 489-509.
3. D. R. Chen, Q. Wu, Y. Ying and D. X. Zhou, Support vector machine soft margin classifiers: error analysis, *J. Mach. Learning Res.*, **5** (2004), 1143-1175.
4. F. Cucker and S. Smale, On the mathematical foundations of learning theory, *Bull. Amer. Math. Soc.*, **39** (2001), 1-49.
5. F. Cucker and D. X. Zhou, *Learing Theory: An Approximation Theory Viewpoint*, Cambridge University Press, 2007.
6. L. Devroye, L. Györfi and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer-Verlag, New York, 1997.
7. Z. Ditzain and V. Totik, *Moduli of Smoothness*, Springer-Verlag, New York, 1987.
8. K. Jetter, J. Stöckler and J. D. Ward, Error estimates for scatterd data interpolation on spheres, *Math. Comput.*, **68** (1999), 733-747.
9. L. Kantorovich, Sur certains developpements suivant les polynomes de la forme de S. Bernstein, I. II, *C. R. Acad. Sci. URSS*, (1930), 563-568, 595-600.
10. I. Steinwart and C. Scovel, Fast rates for support vector machines using Gaussian kernels, *Ann. Statist.*, **35** (2007), 575-607.
11. H. W. Sun and Q. Wu, Least square regression with indefinite kernels and coefficient regularization, *Appl. Comput. Harmonic Anal.*, **30** (2011), 96-109.
12. J. A. K. Suykens and J. Vandewalle, Least squares support vector machine classifiers, *Neural Process. Lett.*, **9** (1999), 293-300.
13. R. Tibshirani, Regression shrinkage and selection via the lasso, *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **58** (1996), 267-288.

14. H. Z. Tong, D. R. Chen and F. H. Yang, Support vector machines regression with l^1-regularizer, *J. Approx. Theory*, **164** (2012), 1331-1344.

15. V. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, New York, 1998.

16. H. Wendland, Local polynomial reproduction and moving least square approximation, *IMA J. Numer. Anal.*, **21** (2001), 285-300.

17. Q. Wu, Y. Ying and D. X. Zhou, Multi-kernel regularized classifiers, *J. Complexity*, **23** (2007), 108-134.

18. Q. Wu and D. X. Zhou, SVM soft margin classifiers: linear programming versus quadratic programming, *Neural Comput.*, **17** (2005), 1160-1187.

19. Q. Wu and D. X. Zhou, Learning with sample dependent hypothesis spaces, *Comput. Math. Appl.*, **56** (2008), 2896-2907.

20. Q. W. Xiao and D. X. Zhou, Learning by nonsymmetric kernel with data dependent spaces and l^1-requlatizer, *Taiwanese J. Math.*, **4** (2010), 1821-1836.

21. T. Zhang, Statistical behavior and consistency of classification methods based on convex risk minimization, *Ann. Statist.*, **32** (2004), 56-85.

22. D. X. Zhou and K. Jetter, Approximation with polynomial kernels and SVM classifiers, *Adv. Comput. Math.*, **25** (2006), 323-344.

Hongzhi Tong
School of Statistics
University of International Business and Economics
Beijing 100029
P. R. China
E-mail: tonghz@uibe.edu.cn

Di-Rong Chen
Department of Mathematics and LMIB
Beijing University of Aeronautics and Astronautics
Beijing 100083
P. R. China
E-mail: drchen@buaa.edu.cn

Fenghong Yang
School of Applied Mathematics
Central University of Finance and Economics
Beijing 100081
P. R. China
E-mail: fhyang@cufe.edu.cn