NOTE **Communicated by Di-Rong Chen**

A Note on Support Vector Machines with Polynomial Kernels

Hongzhi Tong
tonghz@uibe.edu.cn
School of Statistics, University of International Business and Economics, Beijing 100029, P. R. C.

We present a better theoretical foundation of support vector machines with polynomial kernels. The sample error is estimated under Tsybakov's noise assumption. In bounding the approximation error, we take advantage of a geometric noise assumption that was introduced to analyze gaussian kernels. Compared with the previous literature, the error analysis in this note does not require any regularity of the marginal distribution or smoothness of Bayes' rule. We thus establish the learning rates for polynomial kernels for a wide class of distributions.

1 Introduction

Support vector machines (SVMs) as a special kind of kernel based methods were introduced by Boser, Guyon, and Vapnik (1992) with polynomial kernels, and by Cortes and Vapnik (1995) with general kernels. Since then the theoretical foundation of SVMs has been considered by many authors; a far from complete list of papers containing results of this kind includes Chen, Wu, Ying, and Zhou (2004); Cucker and Zhou (2007); Evgeniou, Pontil, and Poggio (2000); Steinwart (2002); Steinwart and Christmann (2008); Steinwart and Scovel (2007); Vapnik (1998); Wu, Ying, and Zhou (2007); Wu and Zhou (2006); Xiang and Zhou (2009); Zhang (2004). In this note, we investigate SVMs classifiers with the polynomial kernels, probably one of the most popular kernels used in SVMs and other kernel-based learning algorithms. Our goal is to establish explicit learning rates for the algorithms under very mild conditions.

We focus on a binary classification problem. Let X be a compact subset of $\mathbb{R}^n$ and $Y = \{-1, 1\}$. A binary classifier is a function $f : X \mapsto Y$, which labels every $\mathbf{x} \in X$ with some $y \in Y$. If ρ is an unknown probability measure on $Z := X \times Y$, then the misclassification error, which is often used to measure the prediction power of a classifier f, can be defined by

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

Here ρ_X is the marginal distribution on X, and $\rho(\cdot|\mathbf{x})$ is the conditional probability measure at $\mathbf{x}$ induced by ρ. If we define the regression function

doi:10.1162/NECO_a_00794

of ρ as

$$f_\rho(\mathbf{x}) := \int_Y y d\rho(y|\mathbf{x}) = \rho(y = 1|\mathbf{x}) - \rho(y = -1|\mathbf{x}), \quad \mathbf{x} \in X,$$

then we know (Devroye, Györfi, & Lugosi, 1997) that the classifier minimizing the misclassification error is called Bayes' rule and is given by $f_c := sgn(f_\rho)$, the sign of f_ρ. Here, for a function $f : X \mapsto \mathbb{R}$, the sign function is defined as $sgn(f)(\mathbf{x}) = 1$ if $f(\mathbf{x}) \geq 0$ and $sgn(f)(\mathbf{x}) = -1$ if $f(\mathbf{x}) < 0$.

As ρ is unknown, f_c cannot be found directly. What we have is a set of samples $\mathbf{z} := \{z_i\}_{i=1}^m = \{(\mathbf{x}_i, y_i)\}_{i=1}^m \in Z^m$ independently drawn according to ρ. Define hinge loss function $V : \mathbb{R} \mapsto \mathbb{R}^+$ as $V(t) = (1-t)_+ = \max\{1-t, 0\}$. Then the SVM classifier is defined as $sgn(f_{\mathbf{z},\lambda})$, where $f_{\mathbf{z},\lambda}$ is a kernel-based regularized empirical risk minimizer—more precisely,

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

where $\mathcal{H}_K$ is a reproducing kernel Hilbert space (RKHS) with Mercer kernel K (see Aronszajn, 1950); $\lambda > 0$ is a regularization parameter, which often depends on m; and $\mathcal{E}_\mathbf{z}(f)$ denotes the empirical risk of a function f associated with hinge loss and samples $\mathbf{z}$, that is,

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

Denote $\|*\|$, $*\cdot*$ the Euclidean norm and inner production of $\mathbb{R}^n$. In most applications two families of kernels are used in equation 1.1. One is gaussian kernel $K_\sigma(\mathbf{x}, \mathbf{y}) = \exp\{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}\}$ with a width parameter $\sigma > 0$. There has been a rich study of SVMs with gaussian kernels in the literature (e.g., Steinwart & Christmann, 2008; Steinwart & Scovel, 2007; Wu et al., 2007; Xiang & Zhou, 2009; Ying & Zhou, 2007). Another important Mercer kernel is the polynomial kernel

$$K_d(\mathbf{x}, \mathbf{y}) = (1 + \mathbf{x} \cdot \mathbf{y})^d, \qquad \mathbf{x}, \mathbf{y} \in \mathbb{R}^n,$$

where d is the degree of the kernel polynomial. It is known (Cucker & Smale, 2001) that the corresponding RKHS $\mathcal{H}_{K_d}$ is the set of n-variable polynomials of degree at most d, and the dimension of $\mathcal{H}_{K_d}$ is $N = \binom{n+d}{d} = \frac{(n+d)!}{n!\,d!}$.

In this note, we restrict our attention to SVM classifiers generated with polynomial kernels; it is defined as $sgn(f_{\mathbf{z},\lambda,d})$, where

$$f_{\mathbf{z},\lambda,d} := \arg\min_{f \in \mathcal{H}_{K_d}} \{\mathcal{E}_\mathbf{z}(f) + \lambda \|f\|_{K_d}^2\}. \tag{1.2}$$

Although the polynomial kernels are the original and probably one of the most important kernels used in SVMs, only a few papers deal with the learning behavior of scheme 1.2. Probably the main difficulty in studying algorithms with polynomial kernels is the approximation error, which involves deeply the degree of kernel polynomial. Zhou and Jetter (2006) partly overcome this difficulty in the univariate case $X = [0, 1]$. An extended version of the result to the multivariate setting is obtained by Tong, Chen, and Peng (2008). However, they require restrictive assumptions on the distribution ρ. For example, it assumes that marginal distribution ρ_X should have a finite distortion with respect to Lebesgue measure on X, and Bayes' rule f_c should satisfy a certain smoothness measured by the modulus of smoothness. These assumptions, however, cannot be easily guaranteed in practice or in theory. In the case of gaussian kernels, a more realistic assumption on ρ is geometric noise assumption (see definition 2), proposed in Steinwart and Scovel (2007), which does not require any kind of smoothness of f_c or any regularity condition on ρ_X. A natural question is whether the geometric noise assumption is still valid to bound the approximation error for polynomial kernels. In this note, we present a positive answer to the question by virtue of the connection of Bernstein polynomial (see definition 4) with probability theory. We then derive the explicit learning rates of SVMs with multivariate polynomial kernels under the geometric noise assumption and the Tsybakov's noise assumption (see definition 1).

2 Definitions, Assumptions and Preliminaries

In this section we introduce some notations, definitions, and basic facts that will be used in this note.

Throughout the note, we assume X to be a simplex on $\mathbb{R}^n$, which is defined by

$$X = \{\mathbf{x} \in \mathbb{R}^n : x_i \geq 0, 1 \leq i \leq n, 1 - |\mathbf{x}| \geq 0\},$$

We use the standard notation: for $\mathbf{x} = (x_1, x_2, \cdots, x_n) \in \mathbb{R}^n$, $|\mathbf{x}| = \sum_{i=1}^n x_i$. We write for $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{k} \in \mathbb{N}^n$ and $d \in \mathbb{N}$,

$$\mathbf{x}^{\mathbf{k}} = x_1^{k_1} x_2^{k_2} \cdots x_n^{k_n}, \quad \mathbf{k}! = k_1! k_2! \cdots k_n!, \text{ and } \binom{d}{\mathbf{k}} = \frac{d!}{\mathbf{k}!(d - |\mathbf{k}|)!}.$$

According to the no-free-lunch theorem (Devroye et al., 1997, theorem 7.2), learning rates are impossible without any restriction on ρ. We shall introduce in this note two kinds of such restrictions. One is the Tsybakov noise condition (see Tsybakov, 2004).

Definition 1. *Let $0 \leq q \leq \infty$. We say that ρ satisfies the Tsybakov noise condition with exponent q if there exists a constant $\hat{c}_q > 0$ such that*

$$\rho_X(\{\mathbf{x} \in X : |f_\rho(\mathbf{x})| \le t\}) \le \hat{c}_q t^q, \quad \forall t > 0. \tag{2.1}$$

It is easy to see that all distributions have at least noise exponent 0. Deterministic distributions (which satisfy $|f_\rho(\mathbf{x})| \equiv 1$) have noise exponent $q = \infty$ with $\hat{c}_\infty = 1$.

Define the generalization error of a function f as

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

The Tsybakov noise condition plays an important role in the possibility of reducing the variance of the relative loss with its expectation. The following lemma will be applied to estimate the sample error in section 4.

Lemma 1 *(Wu and Zhou, 2005). If ρ satisfies equation 2.1, then for every function $f : X \mapsto [-1, 1]$, there exists some constant $c_q > 0$ such that*

$$\mathbb{E}\{[V(yf(\boldsymbol{x})) - V(yf_c(\boldsymbol{x}))]^2\} \le c_q \left(\mathcal{E}(f) - \mathcal{E}(f_c)\right)^{\frac{q}{q+1}}.$$

Another restriction we assume on ρ is the geometric noise condition, introduced in Steinwart and Scovel (2007), in the case of gaussian kernels. To formulate this assumption, we define the classes of X by $X_{-1} := \{\mathbf{x} \in X : f_\rho(\mathbf{x}) < 0\}$, $X_1 := \{\mathbf{x} \in X : f_\rho(\mathbf{x}) > 0\}$, and $X_0 := \{\mathbf{x} \in X : f_\rho(\mathbf{x}) = 0\}$. We also define a distance function $\mathbf{x} \mapsto \tau_\mathbf{x}$ by

$$\tau_\mathbf{x} = \begin{cases} d(\mathbf{x}, X_0 \bigcup X_1), & \text{if} \quad \mathbf{x} \in X_{-1}, \\ d(\mathbf{x}, X_0 \bigcup X_{-1}), & \text{if} \quad \mathbf{x} \in X_1, \\ 0, & \text{otherwise}, \end{cases}$$

where $d(\mathbf{x}, A)$ denotes the distance of $\mathbf{x}$ to a set A with respect to Euclidean norm. With this function, we can define the following geometric noise condition for distributions.

Definition 2. *Let $\alpha > 0$. We say that ρ satisfies the geometric noise condition with exponent α if there exists a constant $c > 0$ such that*

$$\int_X |f_\rho(\boldsymbol{x})| exp\left(-\frac{\tau_x^2}{t}\right) d\rho_X \le ct^\alpha \tag{2.2}$$

holds for all $t > 0$.

The geometric noise condition describes the concentration of the measure $|f_\rho(\mathbf{x})|d\rho_X$ near the decision boundary and does not imply any smoothness

of f_c or regularity of ρ_X with respect to Lebesgue measure on X. However, one can show, as in Steinwart and Scovel (2007, theorem 2.6) that if ρ has a Tsybakov noise exponent q and satisfies the envelope condition $|f_\rho(\mathbf{x})| \leq c_\gamma \tau_{\mathbf{x}}^\gamma$ for some constants γ and c_γ, then ρ has a geometric noise exponent $\alpha = \frac{q+1}{2}\gamma$ if $q \geq 1$, and a geometric noise exponent α for all $\alpha < \frac{q+1}{2}\gamma$ otherwise. (A detailed discussion of this assumption can be found in section 8.2 of Steinwart and Christmann, 2008.)

Geometric noise condition 2.2 will be used in section 3 to estimate the approximation error given just below the proof for proposition 1. To the best of our knowledge, it is the first time that this assumption has been applied to the kernels except for gaussian.

To get better error estimates, one usually makes full use of the projection operator introduced in Chen et al. (2004).

Definition 3. *The projection operator π on the space of functions on X is defined by*

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

Trivially $sgn(\pi(f)) = sgn(f)$. This, together with the comparison theorem proved in Zhang (2004), gives

$$\mathcal{R}(sgn(f)) - \mathcal{R}(f_c) \leq \mathcal{E}(\pi(f)) - \mathcal{E}(f_c), \tag{2.3}$$

for any measurable function $f : X \mapsto \mathbb{R}$. Equation 2.3 asserts that the excess misclassification error can be bounded by means of the excess generalization error, which in turn can be estimated by the following error decomposition technique.

Proposition 1. *Let $f_{\mathbf{z},\lambda,d}$ be defined by equation 1.2. Then for any $f_0 \in \mathcal{H}_{K_d}$,*

$$\mathcal{E}(\pi(f_{\mathbf{z},\lambda,d})) - \mathcal{E}(f_c) \leq A(f_0) + S_{\mathbf{z}}(f_0) - S_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda,d})), \tag{2.4}$$

where $A(f_0) := \mathcal{E}(f_0) - \mathcal{E}(f_c) + \lambda \|f_0\|_{K_d}^2$ and $S_{\mathbf{z}}(f) := (\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}_{\mathbf{z}}(f_c)) - (\mathcal{E}(f) - \mathcal{E}(f_c))$ for any measurable function f.

Proof. Write $\mathcal{E}(\pi(f_{\mathbf{z},\lambda,d})) - \mathcal{E}(f_c)$ as

$$\begin{aligned} &\{(\mathcal{E}(\pi(f_{\mathbf{z},\lambda,d})) - \mathcal{E}(f_c)) - (\mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda,d})) - \mathcal{E}_{\mathbf{z}}(f_c))\} \\ &\quad + \{(\mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda,d})) + \lambda \|f_{\mathbf{z},\lambda,d}\|_{K_d}^2) - (\mathcal{E}_{\mathbf{z}}(f_0) + \lambda \|f_0\|_{K_d}^2)\} \end{aligned}$$

$$+ \{(\mathcal{E}_{\mathbf{z}}(f_0) - \mathcal{E}_{\mathbf{z}}(f_c)) - (\mathcal{E}(f_0) - \mathcal{E}(f_c))\}$$
$$+ \{\mathcal{E}(f_0) - \mathcal{E}(f_c) + \lambda \|f_0\|_{K_d}^2\} - \lambda \|f_{\mathbf{z},\lambda,d}\|_{K_d}^2.$$

It is easy to check that $V(y\pi(f_{\mathbf{z},\lambda,d})(\mathbf{x})) \leq V(yf_{\mathbf{z},\lambda,d}(\mathbf{x}))$; thus, $\mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda,d})) \leq \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda,d})$. This in connection with the definition of $f_{\mathbf{z},\lambda,d}$ implies that the second term is at most zero. The proposition is proved.

The first term and the rest of the terms on the right-hand side of equation 2.4 are called approximation error and sample error (with respect to f_0), respectively.

According to proposition 1, choosing a regularization function $f_0 \in \mathcal{H}_{K_d}$ is key to bounding the excess generalization error. To this end, we introduce the Bernstein polynomials on a simplex (see Lorentz, 1986).

Definition 4. *The Bernstein polynomial for a function f on the simplex X is defined as*

$$B_d(f)(\boldsymbol{x}) := B_{d,n}(f, \boldsymbol{x}) = \sum_{|\boldsymbol{k}| \leq d} f\left(\frac{\boldsymbol{k}}{d}\right) P_{\boldsymbol{k},d}(\boldsymbol{x}), \quad \forall \boldsymbol{x} \in X,$$

where $P_{\boldsymbol{k},d}(\boldsymbol{x}) = \binom{d}{\boldsymbol{k}} \boldsymbol{x}^k (1 - |\boldsymbol{x}|)^{d-|\boldsymbol{k}|}$.

One can easily see that $B_d(f_c) \in \mathcal{H}_{K_d}$. We shall estimate the approximation error and sample error with respect to $B_d(f_c)$ in the next two sections.

3 Approximation Error

In this section we estimate the approximation error $A(B_d(f_c))$: $\mathcal{E}(B_d(f_c)) - \mathcal{E}(f_c) + \lambda \|B_d(f_c)\|_{K_d}^2$. We first compute the RKHS norm $\|B_d(f_c)\|_{K_d}$.

Proposition 2. *If* $|f(\boldsymbol{x})| \leq 1$ *for all* $\boldsymbol{x} \in X$, *then*

$$\|B_d(f)\|_{K_d} < (1 + 4n)^{3d/2}.$$

Proof. Tong et al. (2008, theorem 3.1) showed that

$$\|P_{\mathbf{k},d}\|_{K_d} \leq \left\{ \sum_{|\mathbf{j}| \leq d - |\mathbf{k}|} \binom{d}{\mathbf{k} + \mathbf{j}} 4^{|\mathbf{k}+\mathbf{j}|} \right\}^{1/2}.$$

Since $|f(\mathbf{x})| \le 1$, we have

$$\begin{aligned}
\|B_d(f)\|_{K_d} &\le \sum_{|\mathbf{k}|\le d} \|P_{\mathbf{k},d}\|_{K_d} \\
&\le \sum_{|\mathbf{k}|\le d} \left\{ \sum_{|\mathbf{j}|\le d-|\mathbf{k}|} \binom{d}{\mathbf{k}+\mathbf{j}} 4^{|\mathbf{k}+\mathbf{j}|} \right\}^{1/2} \\
&= \sum_{|\mathbf{k}|\le d} \left\{ \sum_{|\mathbf{k}|\le|\mathbf{i}|\le d} \binom{d}{\mathbf{i}} 4^{|\mathbf{i}|} \right\}^{1/2} \\
&\le \sum_{|\mathbf{k}|\le d} \left\{ \sum_{|\mathbf{i}|\le d} \binom{d}{\mathbf{i}} 4^{|\mathbf{i}|} \right\}^{1/2} \\
&= N\sqrt{(1+4n)^d}.
\end{aligned}$$

Here $N = \binom{n+d}{d}$ is the dimension of $\mathcal{H}_{K_d}$. Note that

$$N = \frac{(n+d)!}{n!\,d!} < (1+4n)^d.$$

The proposition is proved.

Second, we apply the geometric noise condition, equation 2.2, to bound $\mathcal{E}(B_d(f_c)) - \mathcal{E}(f_c)$. To this end, we still need some preparation.

Let $\mathbf{x} = (x_1, x_2, \cdots, x_n) \in X$, $\mathbf{k} \in \mathbb{N}^n$, $d \in \mathbb{N}$ and $|\mathbf{k}| \le d$. Recall that in probability theory, a random vector ξ (in $\mathbb{R}^n$) is said to follow a multinomial distribution with parameters d and $\mathbf{x}$ if it has a probability mass function

$$Prob\{\xi = \mathbf{k}\} = P_{\mathbf{k},d}(\mathbf{x}).$$

Denote $\mathbf{0}$ and $\mathbf{e}_i = (0, \cdots, 0, \overset{i}{1}, 0, \cdots, 0)$, $i = 1, 2, \cdots, n$ the zero vector, and unit vectors of $\mathbb{R}^n$. Let $\{\xi_i\}_{i=1}^d$ be independent and identically distributed random vectors taking values from $\{\mathbf{0}, \mathbf{e}_1, \cdots, \mathbf{e}_n\}$ with probability

$$Prob\{\xi_i = \mathbf{e}_j\} = x_j,\ Prob\{\xi_i = \mathbf{0}\} = 1 - |\mathbf{x}|,\quad i = 1, 2\cdots, d,\ j = 1, 2, \cdots, n.$$

Then one can see that $\eta = \sum_{i=1}^d \xi_i$ follows the multinomial distribution with parameters d and $\mathbf{x}$. We also need the following Bennett inequality for the vector-valued random variable given in Smale and Zhou (2007).

Let $\mathcal{H}$ be a Hilbert space and $\{\zeta_i\}_{i=1}^m$ be m independent random variables with values in $\mathcal{H}$. Suppose that for each i, $\|\zeta_i\|_{\mathcal{H}} < M < \infty$ almost surely. Denote $\sigma^2 = \sum_{i=1}^m \mathbb{E}(\|\zeta_i\|_{\mathcal{H}}^2)$. Then

$$Prob\left\{\left\|\frac{1}{m}\sum_{i=1}^m[\zeta_i - \mathbb{E}\zeta_i]\right\|_{\mathcal{H}} \geq \varepsilon\right\} \leq 2\exp\left\{-\frac{m\varepsilon}{2M}\log\left(1+\frac{mM\varepsilon}{\sigma^2}\right)\right\}, \quad \forall \varepsilon > 0.$$

Applying this probability inequality to $\{\xi_i\}_{i=1}^d$, we find:

Lemma 2. *Let $\boldsymbol{x} \in X$, $d \in \mathbb{N}$. For any $\varepsilon \geq 0$, there holds*

$$\sum_{\|\frac{\mathbf{k}}{d}-\boldsymbol{x}\|\geq\varepsilon} P_{\mathbf{k},d}(\boldsymbol{x}) \leq 2exp\left\{-\frac{d\varepsilon^2}{2(1+\varepsilon)}\right\},$$

where the sum is taken for all values $\mathbf{k} \in \mathbb{N}^n$ satisfying $\left\|\frac{\mathbf{k}}{d} - \boldsymbol{x}\right\| \geq \varepsilon$ and $|\mathbf{k}| \leq d$. Notation of this type is used in the sequel without explanation.

Proof. It is trivial for $\varepsilon = 0$. For $\varepsilon > 0$ we apply the Bennett inequality to $\{\xi_i\}_{i=1}^d$. It satisfies for each $i \in \{1, 2, \cdots, d\}$, $\|\xi_i\| \leq 1$, $\mathbb{E}\xi_i = \mathbf{x}$, $\mathbb{E}(\|\xi_i\|^2) = 1$. Therefore,

$$\begin{aligned}\sum_{\|\frac{\mathbf{k}}{d}-\mathbf{x}\|\geq\varepsilon} P_{\mathbf{k},d}(\mathbf{x}) &= Prob\left\{\left\|\frac{\eta}{d} - \mathbf{x}\right\| \geq \varepsilon\right\} \\ &= Prob\left\{\left\|\frac{1}{d}\sum_{i=1}^d[\xi_i - \mathbb{E}\xi_i]\right\| \geq \varepsilon\right\} \\ &\leq 2\exp\left\{-\frac{d\varepsilon}{2}\log(1+\varepsilon)\right\}.\end{aligned}$$

The lemma thus follows from $\log(1+\varepsilon) \geq \frac{\varepsilon}{1+\varepsilon}$.

Now we can bound $\mathcal{E}(B_d(f_c)) - \mathcal{E}(f_c)$.

Proposition 3. *If ρ satisfies condition 2.2, then*

$$\mathcal{E}(B_d(f_c)) - \mathcal{E}(f_c) \leq 4c5^\alpha d^{-\alpha}.$$

Proof. Since $|f_c(\mathbf{x})| = 1$, it obvious implies $|B_d(f_c)(\mathbf{x})| \leq 1$. Hence the equation

$$\mathcal{E}(f) - \mathcal{E}(f_c) = \int_X |f_\rho(\mathbf{x})||f(\mathbf{x}) - f_c(\mathbf{x})|d\rho_X, \quad f : X \mapsto [-1, 1]$$

given in Zhang (2004) ensures us

$$\mathcal{E}(B_d(f_c)) - \mathcal{E}(f_c) = \int_X |f_\rho(\mathbf{x})||B_d(f_c)(\mathbf{x}) - f_c(\mathbf{x})| d\rho_X. \tag{3.1}$$

In order to estimate $|B_d(f_c)(\mathbf{x}) - f_c(\mathbf{x})|$ for $\mathbf{x} \in X_1$, we observe that $f_c(\mathbf{x}) = 1$ and

$$\begin{aligned}
B_d(f_c)(\mathbf{x}) &= \sum_{|\mathbf{k}|\le d} f_c\left(\frac{\mathbf{k}}{d}\right) P_{\mathbf{k},d}(\mathbf{x}) \\
&= \sum_{|\mathbf{k}|\le d} \left(f_c\left(\frac{\mathbf{k}}{d}\right) + 1\right) P_{\mathbf{k},d}(\mathbf{x}) - 1 \\
&\ge \sum_{\left\|\frac{\mathbf{k}}{d} - \mathbf{x}\right\| < \tau_{\mathbf{x}}} \left(f_c\left(\frac{\mathbf{k}}{d}\right) + 1\right) P_{\mathbf{k},d}(\mathbf{x}) - 1 \\
&= 2 \sum_{\left\|\frac{\mathbf{k}}{d} - \mathbf{x}\right\| < \tau_{\mathbf{x}}} P_{\mathbf{k},d}(\mathbf{x}) - 1 \\
&= 1 - 2 \sum_{\left\|\frac{\mathbf{k}}{d} - \mathbf{x}\right\| \ge \tau_{\mathbf{x}}} P_{\mathbf{k},d}(\mathbf{x}).
\end{aligned}$$

Note that $\tau_{\mathbf{x}} \le \sqrt{2}$ for all $\mathbf{x} \in X$, we can obtain by lemma 2:

$$1 \ge B_d(f_c)(\mathbf{x}) \ge 1 - 4\exp\left(-\frac{d\tau_{\mathbf{x}}^2}{2(1+\sqrt{2})}\right) \ge 1 - 4\exp\left(-\frac{d\tau_{\mathbf{x}}^2}{5}\right)$$

for all $\mathbf{x} \in X_1$. Since for $\mathbf{x} \in X_{-1}$, we can analogously obtain $f_c(\mathbf{x}) = -1$ and

$$-1 \le B_d(f_c)(\mathbf{x}) \le -1 + 4\exp\left(-\frac{d\tau_{\mathbf{x}}^2}{5}\right),$$

we conclude

$$|B_d(f_c)(\mathbf{x}) - f_c(\mathbf{x})| \le 4\exp\left(-\frac{d\tau_{\mathbf{x}}^2}{5}\right) \tag{3.2}$$

for all $\mathbf{x} \in X_1 \bigcup X_{-1}$. When $\mathbf{x} \in X_0$, one has $\tau_{\mathbf{x}} = 0$ and equation 3.2 still holds. Consequently, it follows by equations 3.1, 3.2, and 2.2 with $t = \frac{5}{d}$:

$$\mathcal{E}(B_d(f_c)) - \mathcal{E}(f_c) \le 4\int_X |f_\rho(\mathbf{x})| \exp\left(-\frac{d\tau_{\mathbf{x}}^2}{5}\right) d\rho_X \le 4c\left(\frac{5}{d}\right)^{\alpha}.$$

This proves the proposition.

Combining the results of propositions 2 and 3, we can obtain the estimate of approximation error $A(B_d(f_c))$.

Theorem 1. *Let $d \in \mathbb{N}$, X be the simplex on $\mathbb{R}^n$. If ρ has geometric noise exponent α with constant c in equation 2.2, then there exists a constant $c_\alpha > 0$ depending only on α such that for all $\lambda > 0$,*

$$A(B_d(f_c)) \le c_\alpha((1+4n)^{3d}\lambda + d^{-\alpha}).$$

4 Sample Error

In this section, we estimate the sample error in equation 2.4 with $f_0 = B_d(f_c)$. We first bound $S_{\mathbf{z}}(B_d(f_c))$ by the following one-side Bernstein inequality (see, e.g., Cucker & Smale, 2001; Cucker & Zhou, 2007).

Let ξ be a random variable on a probability space Z with mean μ and variance σ^2. If $|\xi - \mu| \le M$ almost everywhere, then for every $\varepsilon > 0$, there holds

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

Proposition 4. *If ρ satisfies equation 2.1, then for any $0 < \delta < 1$, with the confidence at least $1 - \delta/2$, there holds*

$$S_z(B_d(f_c)) \le \frac{8\log(2/\delta)}{3m} + \left(\frac{2c_q\log(2/\delta)}{m}\right)^{\frac{q+1}{q+2}} + A(B_d(f_c)).$$

Proof. Consider the random variable $\xi(z) = V(yB_d(f_c)(\mathbf{x})) - V(yf_c(\mathbf{x}))$. It satisfies

$$\mu = \mathbb{E}\xi = \mathcal{E}(B_d(f_c)) - \mathcal{E}(f_c); \quad \frac{1}{m}\sum_{i=1}^{m}\xi(z_i) = \mathcal{E}_{\mathbf{z}}(B_d(f_c)) - \mathcal{E}_{\mathbf{z}}(f_c).$$

Since $|f_c(\mathbf{x})| = 1$ and $|B_d(f_c)(\mathbf{x})| \le 1$, one can see $|\xi| \le 2, |\xi - \mu| \le 4$. Applying the one-side Bernstein inequality to ξ yield with at least confidence $1 - \delta/2$,

$$\begin{aligned} S_{\mathbf{z}}(B_d(f_c)) &= (\mathcal{E}_{\mathbf{z}}(B_d(f_c)) - \mathcal{E}_{\mathbf{z}}(f_c)) - (\mathcal{E}(B_d(f_c)) - \mathcal{E}(f_c)) \\ &\le \frac{8\log(2/\delta)}{3m} + \sqrt{\frac{2\sigma^2\log(2/\delta)}{m}}. \end{aligned}$$

Lemma 1 tells us that $\sigma^2 \le \mathrm{E}(\xi^2) \le c_q(\mathcal{E}(B_d(f_c)) - \mathcal{E}(f_c))^{\frac{q}{q+1}}$. This together with Young's inequality implies

$$\begin{aligned}
&\sqrt{\frac{2\sigma^2 \log(2/\delta)}{m}} \\
&\quad \le \sqrt{\frac{2c_q \log(2/\delta)(\mathcal{E}(B_d(f_c)) - \mathcal{E}(f_c))^{\frac{q}{q+1}}}{m}} \\
&\quad \le \frac{q+2}{2(q+1)} \left(\frac{2c_q \log(2/\delta)}{m} \right)^{\frac{q+1}{q+2}} + \frac{q}{2(q+1)} (\mathcal{E}(B_d(f_c)) - \mathcal{E}(f_c)).
\end{aligned}$$

Therefore, with confidence at least $1 - \delta/2$,

$$\begin{aligned}
S_{\mathbf{z}}(B_d(f_c)) &\le \frac{8 \log(2/\delta)}{3m} + \left(\frac{2c_q \log(2/\delta)}{m} \right)^{\frac{q+1}{q+2}} + (\mathcal{E}(B_d(f_c)) - \mathcal{E}(f_c)) \\
&\le \frac{8 \log(2/\delta)}{3m} + \left(\frac{2c_q \log(2/\delta)}{m} \right)^{\frac{q+1}{q+2}} + A(B_d(f_c)).
\end{aligned}$$

Another term of the sample error $-S_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda,d}))$ involves the samples $\mathbf{z}$ and thus runs over a set of functions, so we need some probability inequality for the uniform convergence given by means of the covering numbers.

Definition 5. *For a subset $\mathcal{F}$ of $C(X)$ and $\varepsilon > 0$, the covering number $\mathcal{N}(\mathcal{F}, \varepsilon)$ is defined to be the minimal integer $l \in \mathbb{N}$ such that there exist l balls with radius ε covering $\mathcal{F}$.*

Note that $\mathcal{H}_{K_d} \subset C(X)$. Let $\mathcal{B}_R = \{f \in \mathcal{H}_{K_d} : \|f\|_{K_d} \le R\}$. Tong et al. (2008) showed that

$$\log \mathcal{N}(\mathcal{B}_R, \varepsilon) \le N \log \left(\frac{4 \cdot 2^{d/2} R}{\varepsilon} \right), \tag{4.1}$$

where $N = \binom{n+d}{d}$ is the dimension of $\mathcal{H}_{K_d}$.

The following probability inequality was verified in Wu and Zhou (2005):

Lemma 3. *Let $0 \le \tau \le 1, M > 0$, $B \ge 0$, and $\mathcal{G}$ be a set of functions on Z such that for every $g \in \mathcal{G}$, $\mathbb{E}g \ge 0$, $|g - \mathbb{E}g| \le M$ almost everywhere and $\mathbb{E}(g^2) \le$*

$B(\mathbb{E}g)^\tau$. Then for every $\varepsilon > 0$,

$$Prob_{\mathbf{z}\in Z^m}\left\{\sup_{g\in\mathcal{G}}\frac{\mathbb{E}g-\frac{1}{m}\sum_{i=1}^m g(z_i)}{\sqrt{(\mathbb{E}g)^\tau+\varepsilon^\tau}}>4\varepsilon^{1-\frac{\tau}{2}}\right\}$$
$$\leq \mathcal{N}(\mathcal{G},\varepsilon)exp\left\{-\frac{m\varepsilon^{2-\tau}}{2(B+\frac{1}{3}M\varepsilon^{1-\tau})}\right\}.$$

In order to apply lemma 3 to estimate $-S_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda,d}))$, we need to find a ball $\mathcal{B}_R$ containing $f_{\mathbf{z},\lambda,d}$. The definition of $f_{\mathbf{z},\lambda,d}$ tells us that for $f=0$,

$$\lambda\|f_{\mathbf{z},\lambda,d}\|_{K_d}^2\leq\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\lambda,d})+\lambda\|f_{\mathbf{z},\lambda,d}\|_{K_d}^2\leq\mathcal{E}_{\mathbf{z}}(0)+0=\frac{1}{m}\sum_{i=1}^m V(0)=1.$$

So $\|f_{\mathbf{z},\lambda,d}\|_{K_d}\leq\frac{1}{\sqrt{\lambda}}$. It means that $f_{\mathbf{z},\lambda,d}\in\mathcal{B}_{\frac{1}{\sqrt{\lambda}}}$ for all $\mathbf{z}\in Z^m$ and $\lambda>0$. Applying lemma 3 to the following function set,

$$\mathcal{F}_\lambda:=\left\{V(y\pi(f)(\mathbf{x}))-V(yf_c(\mathbf{x})):f\in\mathcal{B}_{\frac{1}{\sqrt{\lambda}}}\right\},$$

we can find:

Proposition 5. *If ρ satisfies equation 2.1, then for any $0<\delta<1$, with the confidence at least $1-\delta/2$, there holds*

$$-S_z(\pi(f_{\mathbf{z},\lambda,d}))$$
$$\leq\frac{1}{2}(\mathcal{E}(\pi(f_{\mathbf{z},\lambda,d}))-\mathcal{E}(f_c))+20\left\{\left[\frac{4c_q\left(log(2/\delta)+N\,log\left(\frac{2^{d/2}m}{\sqrt{\lambda}}\right)\right)}{m}\right]^{\frac{q+1}{q+2}}\right.$$
$$\left.+\frac{16\left(log(2/\delta)+N\,log\left(\frac{2^{d/2}m}{\sqrt{\lambda}}\right)\right)}{3m}+\frac{4}{m}\right\}.$$

Proof. Each function $g\in\mathcal{F}_\lambda$ has the form $g(z)=V(y\pi(f)(\mathbf{x}))-V(yf_c(\mathbf{x}))$ for some $f\in\mathcal{B}_{\frac{1}{\sqrt{\lambda}}}$. Hence, $\mathbb{E}g=\mathcal{E}(\pi(f))-\mathcal{E}(f_c)\geq 0$ and $\frac{1}{m}\sum_{i=1}^m g(z_i)=\mathcal{E}_{\mathbf{z}}(\pi(f))-\mathcal{E}_{\mathbf{z}}(f_c)$. Since $|\pi(f))(\mathbf{x})|\leq 1$, one has $|g(z)|\leq 2$, $|g-\mathbb{E}g|\leq 4$, and lemma 1 asserts that $\mathbb{E}(g^2)\leq c_q(\mathbb{E}g)^{\frac{q}{q+1}}$. Now applying lemma 3 to $\mathcal{F}_\lambda$,

we can get

$$Prob_{\mathbf{z}\in Z^m}\left\{\sup_{f\in\mathcal{B}_{\frac{1}{\sqrt{\lambda}}}}\frac{(\mathcal{E}(\pi(f))-\mathcal{E}(f_c))-(\mathcal{E}_{\mathbf{z}}(\pi(f))-\mathcal{E}_{\mathbf{z}}(f_c))}{\sqrt{(\mathcal{E}(\pi(f))-\mathcal{E}(f_c))^{\frac{q}{q+1}}+\varepsilon^{\frac{q}{q+1}}}}>4\varepsilon^{\frac{q+2}{2(q+1)}}\right\}$$
$$\leq \mathcal{N}(\mathcal{F}_\lambda,\varepsilon)\exp\left\{-\frac{m\varepsilon^{\frac{q+2}{q+1}}}{2\left(c_q+\frac{4}{3}\varepsilon^{\frac{1}{q+1}}\right)}\right\}.$$

It needs to bound the covering number. Observe that for any $f_1, f_2 \in \mathcal{B}_{\frac{1}{\sqrt{\lambda}}}$, and $(\mathbf{x}, y) \in Z$,

$$\begin{aligned}
&|[V(y\pi(f_1)(\mathbf{x}))-V(yf_c(\mathbf{x}))]-[V(y\pi(f_2)(\mathbf{x}))-V(yf_c(\mathbf{x}))]|\\
&\quad=|V(y\pi(f_1)(\mathbf{x}))-V(y\pi(f_2)(\mathbf{x}))|\\
&\quad\leq |\pi(f_1)(\mathbf{x})-\pi(f_2)(\mathbf{x})|\\
&\quad\leq \|f_1-f_2\|_\infty.
\end{aligned}$$

This in connection with equation 4.1 means that

$$\log\mathcal{N}(\mathcal{F}_\lambda,\varepsilon)\leq\log\mathcal{N}(\mathcal{B}_{\frac{1}{\sqrt{\lambda}}},\varepsilon)\leq N\log\left(\frac{4\cdot 2^{d/2}}{\sqrt{\lambda}\varepsilon}\right).$$

Therefore, if we set ε^* to be the unique positive solution of the equation,

$$N\log\left(\frac{4\cdot 2^{d/2}}{\sqrt{\lambda}\varepsilon}\right)-\frac{m\varepsilon^{\frac{q+2}{q+1}}}{2\left(c_q+\frac{4}{3}\varepsilon^{\frac{1}{q+1}}\right)}=\log(\delta/2).$$

Then with confidence at least $1-\delta/2$,

$$\begin{aligned}
-S_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda,d}))&=(\mathcal{E}(\pi(f_{\mathbf{z},\lambda,d}))-\mathcal{E}(f_c))-(\mathcal{E}_{\mathbf{z}}(\pi(f_{\mathbf{z},\lambda,d}))-\mathcal{E}_{\mathbf{z}}(f_c))\\
&\leq 4\varepsilon^{*\frac{q+2}{2(q+1)}}\sqrt{(\mathcal{E}(\pi(f_{\mathbf{z},\lambda,d}))-\mathcal{E}(f_c))^{\frac{q}{q+1}}+\varepsilon^{*\frac{q}{q+1}}}\\
&\leq 4\varepsilon^{*\frac{q+2}{2(q+1)}}(\mathcal{E}(\pi(f_{\mathbf{z},\lambda,d}))-\mathcal{E}(f_c))^{\frac{q}{2(q+1)}}+4\varepsilon^*\\
&\leq \frac{q}{2(q+1)}(\mathcal{E}(\pi(f_{\mathbf{z},\lambda,d}))-\mathcal{E}(f_c))+\frac{q+2}{2(q+1)}4^{\frac{2(q+1)}{q+2}}\varepsilon^*+4\varepsilon^*\\
&\leq \frac{1}{2}(\mathcal{E}(\pi(f_{\mathbf{z},\lambda,d}))-\mathcal{E}(f_c))+20\varepsilon^*.
\end{aligned}$$

Here the third inequality follows from Young's inequality.

What is left is to bound ε^*. To this end, let

$$h_1(\varepsilon) := \frac{m\varepsilon^{\frac{q+2}{q+1}}}{2\left(c_q + \frac{4}{3}\varepsilon^{\frac{1}{q+1}}\right)}; \quad h_2(\varepsilon) := N\log\left(\frac{4\cdot 2^{d/2}}{\sqrt{\lambda}\varepsilon}\right) - h_1(\varepsilon).$$

One can see that h_2 is a strictly decreasing function on $(0, +\infty)$, and $h_2(\varepsilon^*) = \log(\delta/2)$.

Set

$$s = \left[\frac{4c_q\left(\log(2/\delta) + N\log\left(\frac{2^{d/2}m}{\sqrt{\lambda}}\right)\right)}{m}\right]^{\frac{q+1}{q+2}} + \frac{16\left(\log(2/\delta) + N\log\left(\frac{2^{d/2}m}{\sqrt{\lambda}}\right)\right)}{3m} + \frac{4}{m}.$$

If $\frac{4}{3}s^{\frac{1}{q+1}} \le c_q$, then

$$h_1(s) \ge \frac{ms^{\frac{q+2}{q+1}}}{4c_q} \ge \log(2/\delta) + N\log\left(\frac{2^{d/2}m}{\sqrt{\lambda}}\right).$$

If $\frac{4}{3}s^{\frac{1}{q+1}} > c_q$, then

$$h_1(s) \ge \frac{ms^{\frac{q+2}{q+1}}}{\frac{16}{3}s^{\frac{1}{q+1}}} = \frac{3ms}{16} \ge \log(2/\delta) + N\log\left(\frac{2^{d/2}m}{\sqrt{\lambda}}\right).$$

Thus, in either case we have

$$h_1(s) \ge \log(2/\delta) + N\log\left(\frac{2^{d/2}m}{\sqrt{\lambda}}\right).$$

On the other hand, since $s > \frac{4}{m}$, it follows that

$$h_2(s) \le N\log\left(\frac{2^{d/2}m}{\sqrt{\lambda}}\right) - h_1(s) \le \log(\delta/2) = h_2(\varepsilon^*).$$

Therefore, $\varepsilon^* \le s$. The proof of the proposition is complete.

5 Learning Rates

In this section we derive explicit learning rates for equation 1.2 by appropriately choosing the regularization parameter $\lambda = \lambda(m)$ and the degree $d = d(m)$ of the kernel polynomial.

Theorem 2. *Assume that ρ satisfies equations 2.1 and 2.2. Let $\theta = \frac{q+1}{(n+1)(q+1)+\alpha(q+2)}$, $d = m^\theta$, $\lambda = exp\{-3(n+2)m^\theta\}$. Then for all $0 < \delta < 1$, with confidence at least $1 - \delta$, we have*

$$\mathcal{R}(sgn(f_{\mathbf{z},\lambda,d})) - \mathcal{R}(f_c) \le C log(2/\delta + 1)m^{-\theta\alpha}, \tag{5.1}$$

where C is some constant independent of m or δ.

Proof. Putting theorem 1 and propositions 4 and 5 into proposition 1 with $f_0 = B_d(f_c)$, we can find that with confidence at least $1 - \delta$,

$$\begin{aligned}
&\mathcal{E}(\pi(f_{\mathbf{z},\lambda,d})) - \mathcal{E}(f_c) \\
&\quad \le \frac{1}{2}(\mathcal{E}(\pi(f_{\mathbf{z},\lambda,d})) - \mathcal{E}(f_c)) + 2c_\alpha((1+4n)^{3d}\lambda + d^{-\alpha}) + \frac{8\log(2/\delta)}{3m} \\
&\qquad + \left(\frac{2c_q \log(2/\delta)}{m}\right)^{\frac{q+1}{q+2}} + 20\left\{\left[\frac{4c_q\left(\log(2/\delta) + N\log\left(\frac{2^{d/2}m}{\sqrt{\lambda}}\right)\right)}{m}\right]^{\frac{q+1}{q+2}}\right. \\
&\qquad \left. + \frac{16\left(\log(2/\delta) + N\log\left(\frac{2^{d/2}m}{\sqrt{\lambda}}\right)\right)}{3m} + \frac{4}{m}\right\} \\
&\quad \le \frac{1}{2}(\mathcal{E}(\pi(f_{\mathbf{z},\lambda,d})) - \mathcal{E}(f_c)) + (\log(2/\delta + 1))\left(190 + 21(8c_q)^{\frac{q+1}{q+2}}\right)m^{-\frac{q+1}{q+2}} \\
&\qquad + 2c_\alpha((1+4n)^{3d}\lambda + d^{-\alpha}) \\
&\qquad + 20\left\{\left[\frac{8c_q N\log\left(\frac{2^{d/2}m}{\sqrt{\lambda}}\right)}{m}\right]^{\frac{q+1}{q+2}} + \frac{16N\log\left(\frac{2^{d/2}m}{\sqrt{\lambda}}\right)}{3m}\right\}.
\end{aligned}$$

Since $N = \frac{(n+d)!}{n!d!} \le (2d)^n$, by taking $\theta > 0$, $d = m^\theta$, $\lambda = \exp\{-3(n+2)m^\theta\}$, we have with the same confidence,

$$\begin{aligned}
&\mathcal{E}(\pi(f_{\mathbf{z},\lambda,d})) - \mathcal{E}(f_c) \\
&\quad \le 2(\log(2/\delta+1)) \left(190 + 21(8c_q)^{\frac{q+1}{q+2}}\right) m^{-\frac{q+1}{q+2}} \\
&\qquad + 4c_\alpha \left(\left(\frac{1+4n}{e^{n+2}}\right)^{3m^\theta} + m^{-\alpha\theta}\right) \\
&\qquad + 40 \left\{\left[2^{n+3} c_q m^{n\theta-1} \left(\left(\frac{\log 2}{2} + n + 2\right) m^\theta + \log m\right)\right]^{\frac{q+1}{q+2}}\right. \\
&\qquad \left. + 2^{n+3} m^{n\theta-1} \left(\left(\frac{\log 2}{2} + n + 2\right) m^\theta + \log m\right)\right\}.
\end{aligned}$$

It is easy to see $\left(\frac{(1+4n)}{e^{n+2}}\right)^{3m^\theta} \le e^{-3m^\theta}$. Recall an elementary inequality

$$e^{-cx} \le \left(\frac{a}{ec}\right)^a x^{-a} \quad \forall x,\ a,\ c > 0. \tag{5.2}$$

This inequality is verified by considering the function $f(x) = x^a e^{-cx}$, which is maximized at $x = \frac{a}{c}$. Applying equation 5.2 with $x = m^\theta,\ c = 3,\ a = \alpha$, we have

$$e^{-3m^\theta} \le \left(\frac{\alpha}{3e}\right)^\alpha m^{-\alpha\theta}.$$

Applying equation 5.2 with $x = \log m,\ c = \theta,\ a = 1$ again, we have

$$\log m \le \frac{1}{\theta e} m^\theta.$$

Therefore, with confidence at least $1 - \delta$,

$$\begin{aligned}
&\mathcal{E}(\pi(f_{\mathbf{z},\lambda,\lambda})) - \mathcal{E}(f_c) \\
&\quad \le \widetilde{C} \log(2/\delta+1) \left(m^{-\frac{q+1}{q+2}} + m^{-\alpha\theta} + m^{\frac{((n+1)\theta-1)(q+1)}{q+2}} + m^{(n+1)\theta-1}\right),
\end{aligned}$$

where

$$\begin{aligned}
\widetilde{C} := &\, 2\left(190 + 21(8c_q)^{\frac{q+1}{q+2}}\right) + 4c_\alpha \left(\left(\frac{\alpha}{3e}\right)^\alpha + 1\right) \\
&+ 40 \left\{\left[2^{n+3} c_q \left(\frac{\log 2}{2} + n + 2 + \frac{1}{\theta e}\right)\right]^{\frac{q+1}{q+2}}\right. \\
&\left. + 2^{n+3} \left(\frac{\log 2}{2} + n + 2 + \frac{1}{\theta e}\right)\right\}.
\end{aligned}$$

By taking $\theta = \frac{q+1}{(n+1)(q+1)+\alpha(q+2)}$, the conclusion then follows from equation 2.3 and $C = 4\widetilde{C}$.

When the Tsybakov noise condition, equation 2.1, is not assumed, one can still use theorem 2 by setting $q = 0$ and obtain learning rate $m^{-\frac{\alpha}{n+1+2\alpha}}$. When q tends to infinity, the power index of m in equation 5.1 has the limit $-\frac{\alpha}{n+1+\alpha}$, which can be very close to -1 for large α. So the learning rate in theorem 2 can be $m^{\epsilon-1}$ for arbitrarily small $\epsilon > 0$ when q and α are large enough.

Acknowledgments

I thank Qiang Wu for his helpful discussions. This work was completed when I was a visiting scholar at Middle Tennessee State University. It is partly supported by NSF of China under grant 11501380 and the Fundamental Research Funds for the Central Universities in UIBE (13YBLG01).

References

Aronszajn, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, *68*, 337–404.

Boser, B. E., Guyon, I., & Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop of Computational Learning Theory*, *5* (pp. 144–152). New York: ACM.

Chen, D. R., Wu, Q., Ying, Y., & Zhou, D. X. (2004). Support vector machine soft margin classifiers: Error analysis. *J. Mach. Learn. Res.*, *5*, 1143–1175.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Mach. Learning*, *20*, 273–297.

Cucker, F., & Smale, S. (2001). On the mathematical foundations of learning theory. *Bull. Amer. Math. Soc.*, *39*, 1–49.

Cucker, F., & Zhou, D. X. (2007). *Learning theory: An approximation theory viewpoint*. Cambridge: Cambridge University Press.

Devroye, L., Györfi, L., & Lugosi, G. (1997). *A probabilistic theory of pattern recognition*. New York: Springer-Verlag.

Evgeniou, T., Pontil, M., & Poggio, T. (2000). Regularization networks and support vector machines. *Adv. Comput. Math.*, *13*, 1–50.

Lorentz, G. G. (1986). *Bernstein polynomials*. New York: Chelsea.

Smale, S., & Zhou, D. X. (2007). Learning theory estimates via integral operators and their applications. *Constr. Approx.*, *26*, 153–172.

Steinwart, I. (2002). Support vector machines are universally consistent. *J. Complexity*, *18*, 768–791.

Steinwart, I., & Christmann, A. (2008). *Support vector machines*. New York: Springer.

Steinwart, I., & Scovel, C. (2007). Fast rates for support vector machines using gaussian kernels. *Ann. Statist.*, *35*, 575–607.

Tong, H. Z., Chen, D. R., & Peng, L. Z. (2008). Learning rates for regularized classifiers using multivariate polynomial kernels. *J. Complexity.*, *24*, 619–631.

Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Stat.*, *32*, 135–166.

Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley.

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

Wu, Q., & Zhou, D. X. (2005). SVM soft margin classifiers: Linear programming versus quadratic programming. *Neural Comput.*, *17*, 1160–1187.

Wu, Q.& Zhou, D. X. (2006). Analysis of support vector machine classification. *J. Comput. Anal. Appl.*, *8*, 99–119.

Xiang, D. H., & Zhou, D. X. (2009). Classification with gaussian and convex loss. *J. Mach. Learn. Res.*, *10*, 1447–1468.

Ying, Y., & Zhou, D. X. (2007). Learnability of gaussians with flexible variances. *J. Mach. Learn. Res.*, *8*, 249–276.

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

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

Received July 27, 2015; accepted September 2, 2015.