Learning Rates of Least-square Regularized Regression

Qiang Wu, Yiming Ying and Ding-Xuan Zhou†

Department of Mathematics, City University of Hong Kong
Kowloon, Hong Kong, CHINA
wu.qiang@student.cityu.edu.hk, ymying@cityu.edu.hk, mazhou@math.cityu.edu.hk

Abstract

This paper considers the regularized learning algorithm associated with the least-square loss and reproducing kernel Hilbert spaces. The target is the error analysis for the regression problem in learning theory. A novel regularization approach is presented, which yields satisfactory learning rates. The rates depend on the approximation property and the capacity of the reproducing kernel Hilbert space measured by covering numbers. When the kernel is C^∞ and the regression function lies in the corresponding reproducing kernel Hilbert space, the rate is $m^{-\zeta}$ with ζ arbitrarily close to 1, regardless of the variance of the bounded probability distribution.

Short Title: Least-square Regularized Regression

Keywords and Phrases: learning theory, reproducing kernel Hilbert space, regularization error, covering number, regularization scheme.

AMS Subject Classification Numbers: 68T05, 62J02.

† Corresponding author: Ding-Xuan Zhou. Tel: (852) 2788 9708; Fax: (852) 2788 8561.

§1. Introduction

In this paper we consider the least-square regularized algorithm for the regression problem. The main results will be satisfactory learning rates.

Let (X, d) be a compact metric space and $Y = \mathbb{R}$. Let ρ be a probability distribution on $Z := X \times Y$. The *error* (or generalization error) for a function $f : X \to Y$ is defined as

$$\mathcal{E}(f) := \int_Z \big(f(x) - y\big)^2 d\rho. \tag{1.1}$$

The function that minimizes the error is called the *regression function*. It is given by

$$f_\rho(x) = \int_Y y d\rho(y|x), \qquad x \in X \tag{1.2}$$

where $\rho(\cdot|x)$ is the conditional probability measure at x induced by ρ.

The target of the regression problem is to learn the regression function or find good approximations from random samples.

The least-square regularized algorithm for the regression problem is a discrete least-square problem associated with a Mercer kernel.

Let $K : X \times X \to \mathbb{R}$ be continuous, symmetric and positive semidefinite, *i.e.*, for any finite set of distinct points $\{x_1, \cdots, x_\ell\} \subset X$, the matrix $(K(x_i, x_j))_{i,j=1}^\ell$ is positive semidefinite. Such a function is called a *Mercer kernel*.

The *Reproducing Kernel Hilbert Space* (RKHS) $\mathcal{H}_K$ associated with the kernel K is defined (see [3]) to be the closure of the linear span of the set of functions $\{K_x := K(x, \cdot) : x \in X\}$ with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_K} = \langle \cdot, \cdot \rangle_K$ satisfying $\langle K_x, K_y \rangle_K = K(x, y)$. That is, $\langle \sum_i \alpha_i K_{x_i}, \sum_j \beta_j K_{y_j} \rangle_K = \sum_{i,j} \alpha_i \beta_j K(x_i, y_j)$. The reproducing property takes the form

$$\langle K_x, f \rangle_K = f(x), \qquad \forall\ x \in X,\ f \in \mathcal{H}_K.$$

Denote $C(X)$ as the space of continuous functions on X with the norm $\| \cdot \|_\infty$. Let $\kappa = \sup_{x \in X} \sqrt{K(x, x)}$. Then the above reproducing property tells us that

$$\|f\|_\infty \le \kappa \|f\|_K, \qquad \forall\ f \in \mathcal{H}_K. \tag{1.3}$$

Now the *least-square regularized algorithm* for the regression problem associated with the Mercer kernel K is defined to be the minimizer of the following least-square optimization problem involving a set of random samples $\mathbf{z} := \{z_i\}_{i=1}^m = \{(x_i, y_i)\}_{i=1}^m \in Z^m$ independently drawn according to ρ:

$$f_{\mathbf{z}} = f_{\mathbf{z},\lambda} := \arg\min_{f\in\mathcal{H}_K} \left\{ \frac{1}{m}\sum_{i=1}^m \big(f(x_i) - y_i\big)^2 + \lambda\|f\|_K^2 \right\}. \tag{1.4}$$

Here $\lambda \geq 0$ is a constant called the *regularization parameter*. Usually it is chosen to depend on m: $\lambda = \lambda(m)$, and $\lim_{m\to\infty} \lambda(m) = 0$.

Throughout this paper, we assume that for some $M \geq 0$, $\rho(\cdot|x)$ is almost everywhere supported on $[-M, M]$, that is, $|y| \leq M$ almost surely (with respect to ρ). It follows from the definition (1.2) of f_ρ that $|f_\rho(x)| \leq M$.

The efficiency of the algorithm (1.4) is measured by the difference between $f_{\mathbf{z}}$ and the regression function f_ρ. Because of the least-square nature, the measurement is the weighted L^2 metric in $L^2_{\rho_X}$ defined as $\|f\|_\rho = \|f\|_{L^2_{\rho_X}} := \left(\int_X |f(x)|^2 d\rho_X\right)^{1/2}$, where ρ_X is the marginal distribution of ρ on X. One can see [8] that

$$\|f_{\mathbf{z}} - f_\rho\|_\rho^2 = \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho). \tag{1.5}$$

Estimating the error (1.5) for the least-square regression algorithm (1.4) by means of properties of ρ and K is our goal. In particular, we shall show how the choice of the regularization parameter $\lambda = \lambda(m)$ in the algorithm affects the learning rates.

Set the empirical error as

$$\mathcal{E}_{\mathbf{z}}(f) := \frac{1}{m}\sum_{i=1}^m \big(f(x_i) - y_i\big)^2.$$

It is a discretization of the error $\mathcal{E}(f)$. Since the scheme (1.4) can be written as

$$f_{\mathbf{z}} = \arg\min_{f\in\mathcal{H}_K} \left\{\mathcal{E}_{\mathbf{z}}(f) + \lambda\|f\|_K^2\right\}, \tag{1.6}$$

we would expect that the minimizer of the regularized empirical error, $f_{\mathbf{z}}$, is a good approximation of the minimizer f_ρ of the error $\mathcal{E}(f)$, as $m \to \infty$ and $\lambda = \lambda(m) \to 0$. This

is actually true if f_ρ can be approximated by functions from $\mathcal{H}_K$, measured by the decay (to 0 as λ becomes small) of the *regularization error of the scheme* (1.6) defined as

$$\mathcal{D}(\lambda) := \inf_{f \in \mathcal{H}_K} \left\{ \|f - f_\rho\|_\rho^2 + \lambda \|f\|_K^2 \right\}. \tag{1.7}$$

Since the minimization (1.6) is taken for the discrete quantity $\mathcal{E}_{\mathbf{z}}$, the approximation of f_ρ by $f_{\mathbf{z}}$ involves the capacity of the function space $\mathcal{H}_K$. Here the capacity is measured by the covering number of the balls B_R (considered as a subset of $C(X)$)

$$B_R := \{f \in \mathcal{H}_K : \|f\|_K \le R\}.$$

Definition 1. *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 $\ell \in \mathbb{N}$ such that there exist ℓ disks with radius η covering $\mathcal{F}$.*

Denote the covering number of B_1 in $C(X)$ with the metric $\|\cdot\|_\infty$ by $\mathcal{N}(\eta)$.

Definition 2. *We say that the Mercer kernel K has polynomial complexity exponent $s > 0$ if*

$$\log \mathcal{N}(\eta) \le C_0 \left(1/\eta\right)^s, \qquad \forall \eta > 0. \tag{1.8}$$

The covering number $\mathcal{N}(\eta)$ has been extensively studied, see e.g. [5, 29, 30]. It was shown in [30] that (1.8) holds if K is $C^{2n/s}$ on a subset X of $\mathbb{R}^n$. In particular, for a C^∞ kernel (such as Gaussians), (1.8) is valid for any $s > 0$.

Let us state our main result on the error analysis.

Theorem 1. *Let $f_{\mathbf{z}}$ be defined by (1.4). Assume (1.8) and that f_ρ is not identically zero. For any $0 < \zeta < \frac{1}{1+s}$, $0 < \delta < 1$ and $m \ge m_{\delta,\zeta}$, with confidence $1 - \delta$ we have*

$$\|f_{\mathbf{z}} - f_\rho\|_\rho^2 \le \frac{\widetilde{C}}{\|f_\rho\|_\rho^2} \log\left(\frac{2}{\delta}\right) \mathcal{D}\left(m^{-\zeta}\right), \qquad \text{by taking } \lambda = m^{-\zeta} \tag{1.9}$$

where $m_{\delta,\zeta}$ and $\widetilde{C}$ are constants depending on C_0, s, ζ, κ, M, and $m_{\delta,\zeta}$ also on δ.

The above two constants $m_{\delta,\zeta}$ and $\widetilde{C}$ will be explicitly given in the proof of Theorem 1 in Section 5. The assumption that f_ρ is not identically zero is necessary: otherwise

$f_\rho \equiv 0$ would imply $\mathcal{D}(\lambda) \equiv 0$ which cannot bound $\|f_{\mathbf{z}} - f_\rho\|_\rho^2$ (a quantity depending on the variance of ρ).

The bound in (1.9) contains the quantity $\|f_\rho\|_\rho^2$ in the denominator of the constant term. It cannot be removed as shown by an example in Section 5. However, a modified bound, without $\|f_\rho\|_\rho^2$, with $\frac{\widetilde{C}}{\|f_\rho\|_\rho^2}\mathcal{D}\left(m^{-\zeta}\right)$ replaced by $\widetilde{C}\left(\mathcal{D}\left(m^{-\zeta}\right) + 2m^{-\frac{1}{1+s}}\right)$ will be illustrated by Theorem 2 in Section 5.

When $s < 1$, the learning rates derived from Theorem 1 are better than those in the literature of regularization schemes where the best kernel independent learning rate is $\left(\frac{1}{m}\right)^{\frac{1}{2}}$. For detailed comparisons, see Section 6. The rates are achieved by an iteration technique which may be useful for studying other algorithms in learning theory. The following example is a special case of Corollary 6.2 with $r = 1/2$.

Proposition 1.1. *Let $f_{\mathbf{z}}$ be defined by (1.4). Assume K is C^∞ on $X \subset \mathbb{R}^n$ and $f_\rho \in \mathcal{H}_K$. Take $\lambda = \lambda(m) = m^{2\epsilon-1}$ with $\epsilon > 0$. Then for any $0 < \delta < 1$, with confidence $1 - \delta$*

$$\|f_{\mathbf{z}} - f_\rho\|_\rho^2 \le \widetilde{C}\log(\frac{2}{\delta})\left(\frac{1}{m}\right)^{1-\epsilon}$$

for $m \ge m_{\delta,\epsilon}$, where the constants $m_{\delta,\epsilon}$ $\widetilde{C}$ are independent of m.

For the error analysis, we use a regularization approach which we introduced in [25].

To estimate $\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho)$, we introduce a *regularizing function* $\widetilde{f}_\lambda \in \mathcal{H}_K$. It is arbitrarily chosen and depends on λ. A standard choice is

$$f_\lambda := \arg\min_{f\in\mathcal{H}_K}\left\{\|f - f_\rho\|_\rho^2 + \lambda\|f\|_K^2\right\}. \tag{1.10}$$

Proposition 1.2. *Let $\widetilde{f}_\lambda \in \mathcal{H}_K$, and $f_{\mathbf{z}}$ be defined by (1.6). Then $\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) \le \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) + \lambda\|f_{\mathbf{z}}\|_K^2$ which can be bounded by*

$$\left\{\mathcal{E}(\widetilde{f}_\lambda) - \mathcal{E}(f_\rho) + \lambda\|\widetilde{f}_\lambda\|_K^2\right\} + \left\{\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \mathcal{E}_{\mathbf{z}}(\widetilde{f}_\lambda) - \mathcal{E}(\widetilde{f}_\lambda)\right\}. \tag{1.11}$$

Proof. Write $\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) + \lambda\|f_{\mathbf{z}}\|_K^2$ as

$$\begin{aligned}&\left\{\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}})\right\} + \left\{\left(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \lambda\|f_{\mathbf{z}}\|_K^2\right) - \left(\mathcal{E}_{\mathbf{z}}(\widetilde{f}_\lambda) + \lambda\|\widetilde{f}_\lambda\|_K^2\right)\right\}\\ &\quad + \left\{\mathcal{E}_{\mathbf{z}}(\widetilde{f}_\lambda) - \mathcal{E}(\widetilde{f}_\lambda)\right\} + \left\{\mathcal{E}(\widetilde{f}_\lambda) - \mathcal{E}(f_\rho) + \lambda\|\widetilde{f}_\lambda\|_K^2\right\}.\end{aligned}$$

The expression (1.6) tells us that the second term is at most zero since $\widetilde{f}_\lambda \in \mathcal{H}_K$. Hence $\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) + \lambda \|f_{\mathbf{z}}\|_K^2$ is bounded by (1.11). □

The first term in (1.11) is called the regularization error [19]. It can be expressed as a K-functional, since for any measurable function $f : X \to \mathbb{R}$, there holds

$$\|f - f_\rho\|_\rho^2 = \mathcal{E}(f) - \mathcal{E}(f_\rho). \tag{1.12}$$

Definition 3. *The regularization error for a regularizing function $\widetilde{f}_\lambda \in \mathcal{H}_K$ is defined as*

$$\widetilde{\mathcal{D}}(\lambda) := \mathcal{E}(\widetilde{f}_\lambda) - \mathcal{E}(f_\rho) + \lambda \|\widetilde{f}_\lambda\|_K^2. \tag{1.13}$$

According to (1.12), the regularization error $\widetilde{\mathcal{D}}(\lambda)$ for the special regularizing function f_λ becomes the regularization error $\mathcal{D}(\lambda)$ of the scheme (1.6), defined by (1.7).

The regularization error for the least-square error is well understood [20, 18]. The rate of the regularization error is not only important for bounding the first term in (1.11), but also crucial for bounding the second term called the *sample error*. The decay of $\lambda = \lambda(m)$ (as $m \to \infty$) determines the size of the hypothesis space and hence the sample error estimates. Therefore, we need to understand the choice of the parameter λ from the bound for $\widetilde{\mathcal{D}}(\lambda)$.

Write the sample error in (1.11) as

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \mathcal{E}_{\mathbf{z}}(\widetilde{f}_\lambda) - \mathcal{E}(\widetilde{f}_\lambda) = \left\{ E(\xi_1) - \frac{1}{m}\sum_{i=1}^m \xi_1(z_i) \right\} + \left\{ \frac{1}{m}\sum_{i=1}^m \xi_2(z_i) - E(\xi_2) \right\}, \tag{1.14}$$

where

$$\xi_1 := \big(f_{\mathbf{z}}(x) - y\big)^2 - \big(f_\rho(x) - y\big)^2, \qquad \xi_2 := \big(\widetilde{f}_\lambda(x) - y\big)^2 - \big(f_\rho(x) - y\big)^2. \tag{1.15}$$

In (1.15), ξ_2 is a fixed random variable on (Z, ρ) with mean $E(\xi_2) = \mathcal{E}(f_\lambda) - \mathcal{E}(f_\rho)$. But ξ_1 is not a single random variable on Z, it depends not only on the variable $z \in Z$, but also on the sample $\mathbf{z}$. We have abused the notion $E(\xi_1) = \int_Z \xi_1(x, y) d\rho = \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho)$.

The last term of (1.14) is a typical quantity that can be estimated by probability inequalities. We shall bound this term by a Bernstein inequality in Section 2.

The function $f_{\mathbf{z}}$ changing with the sample $\mathbf{z}$ runs over a set of functions, and should not be considered as a fixed function. In Section 3 we shall bound the sample error part involving ξ_1 in (1.14). To this end, we shall use the covering number $\mathcal{N}(\eta)$ of the unit ball B_1 of $\mathcal{H}_K$.

§2. Part of the Sample Error Involving the Regularizing Function

In this section we bound the last term of (1.14): $\frac{1}{m}\sum_{i=1}^m \xi_2(z_i) - E(\xi_2)$. Here we apply the one-side Bernstein inequality:

Let ξ be a random variable on a probability space Z with mean $E(\xi) = \mu$, variance $\sigma^2(\xi) = \sigma^2$, and satisfying $|\xi(z) - E(\xi)| \le M_\xi$ for almost all $z \in Z$. Then for all $\varepsilon > 0$,

$$\mathrm{Prob}_{\mathbf{z}\in Z^m}\left\{\frac{1}{m}\sum_{i=1}^m \xi(z_i) - \mu \ge \varepsilon\right\} \le \exp\left\{-\frac{m\varepsilon^2}{2\left(\sigma^2 + \frac{1}{3}M_\xi\varepsilon\right)}\right\}.$$

Proposition 2.1. *For every $0 < \delta < 1$, with confidence at least $1-\delta$, there holds*

$$\frac{1}{m}\sum_{i=1}^m \xi_2(z_i) - E(\xi_2) \le \widetilde{\mathcal{D}}(\lambda)\left(1 + \frac{4\kappa^2\log\big(1/\delta\big)}{m\lambda}\right) + \frac{36M^2\log\big(1/\delta\big)}{m}.$$

Proof. From the definition of $\widetilde{\mathcal{D}}(\lambda)$, we know that

$$\lambda\|\widetilde{f}_\lambda\|_K^2 \le \mathcal{E}(\widetilde{f}_\lambda) - \mathcal{E}(f_\rho) + \lambda\|\widetilde{f}_\lambda\|_K^2 = \widetilde{\mathcal{D}}(\lambda).$$

It follows from (1.3) that

$$\|\widetilde{f}_\lambda\|_\infty \le \kappa\|\widetilde{f}_\lambda\|_K \le \kappa\sqrt{\widetilde{\mathcal{D}}(\lambda)/\lambda}.$$

Observe that

$$\xi_2 = \big(\widetilde{f}_\lambda(x) - f_\rho(x)\big)\big\{\big(\widetilde{f}_\lambda(x) - y\big) + \big(f_\rho(x) - y\big)\big\}.$$

Since $|f_\rho(x)| \le M$ almost everywhere, we have

$$|\xi_2| \le \big(\|\widetilde{f}_\lambda\|_\infty + M\big)\big(\|\widetilde{f}_\lambda\|_\infty + 3M\big) \le c := \big(\kappa\sqrt{\widetilde{\mathcal{D}}(\lambda)/\lambda} + 3M\big)^2.$$

Hence $|\xi_2(z) - E(\xi_2)| \le M_{\xi_2} := 2c$. Moreover, $E(\xi_2^2)$ equals

$$E\left(\big(\widetilde{f}_\lambda(x) - f_\rho(x)\big)^2\big\{\big(\widetilde{f}_\lambda(x) - y\big) + \big(f_\rho(x) - y\big)\big\}^2\right) \le \|\widetilde{f}_\lambda - f_\rho\|_\rho^2\big(\|\widetilde{f}_\lambda\|_\infty + 3M\big)^2$$

which implies that $\sigma^2(\xi_2) \le E(\xi_2^2) \le c\widetilde{\mathcal{D}}(\lambda)$.

Now we apply the one-side Bernstein inequality to ξ_2. It asserts that for any $t > 0$,

$$\frac{1}{m}\sum_{i=1}^{m}\xi_2(z_i) - E(\xi_2) \le t$$

with confidence at least

$$1 - \exp\left\{-\frac{mt^2}{2\big(\sigma^2(\xi_2) + \frac{1}{3}M_{\xi_2}t\big)}\right\} \ge 1 - \exp\left\{-\frac{mt^2}{2c\big(\widetilde{\mathcal{D}}(\lambda) + \frac{2}{3}t\big)}\right\}.$$

Choose t^* to be the unique positive solution of the quadratic equation

$$-\frac{mt^2}{2c\big(\widetilde{\mathcal{D}}(\lambda) + \frac{2}{3}t\big)} = \log\delta.$$

Then with confidence $1-\delta$, there holds $\frac{1}{m}\sum_{i=1}^{m}\xi_2(z_i) - E(\xi_2) \le t^*$. But

$$\begin{aligned} t^* &= \left(\frac{2c}{3}\log\big(1/\delta\big) + \sqrt{\Big(\frac{2c}{3}\log\big(1/\delta\big)\Big)^2 + 2cm\log\big(1/\delta\big)\widetilde{\mathcal{D}}(\lambda)}\right)/m \\ &\le \frac{4c\log\big(1/\delta\big)}{3m} + \sqrt{2c\log\big(1/\delta\big)\widetilde{\mathcal{D}}(\lambda)/m} \le \frac{4c\log\big(1/\delta\big)}{3m} + \widetilde{\mathcal{D}}(\lambda) + \frac{c\log\big(1/\delta\big)}{2m}. \end{aligned}$$

Recall $c = (\kappa\sqrt{\widetilde{\mathcal{D}}(\lambda)/\lambda} + 3M)^2$. It follows that

$$t^* \le \widetilde{\mathcal{D}}(\lambda)\left(1 + \frac{4\kappa^2\log\big(1/\delta\big)}{m\lambda}\right) + \frac{36M^2\log\big(1/\delta\big)}{m}.$$

This implies the desired estimate. □

§3. Part of the Sample Error Involving $f_{\mathbf{z}}$

In this section we estimate the first part of (1.14). It is more difficult to deal with because ξ_1 involves the sample $\mathbf{z}$ through $f_{\mathbf{z}}$. We will use the idea of empirical risk minimization [22, 12, 5, 8, 23] to bound this term by means of a covering number.

The following ratio probability inequality (stated in Lemmas 3.1 and 3.2) is a standard result in learning theory (e.g. [14, 8, 26]). It deals with variances for a function class, since the Bernstein inequality takes care of the variance well only for a single random variable (see Proposition 2.1). Our current form was motivated by sample error estimates for the square loss [4, 5, 8, 14, 7].

Lemma 3.1. *Suppose a random variable ξ on Z satisfies $\mu = E(\xi) \geq 0$, and $\mathbf{z} = (z_i)_{i=1}^m$ are independent samples. If $|\xi - \mu| \leq B$ almost everywhere and $E(\xi^2) \leq c_\xi E(\xi)$ for some $c_\xi \geq 0$, then for every $\varepsilon > 0$ and $0 < \alpha \leq 1$, there holds*

$$\mathrm{Prob}_{\mathbf{z}\in Z^m}\left\{\frac{\mu - \frac{1}{m}\sum_{i=1}^m \xi(z_i)}{\sqrt{\mu+\varepsilon}} \geq \alpha\sqrt{\varepsilon}\right\} \leq \exp\left\{-\frac{\alpha^2 m\varepsilon}{2c_\xi + \frac{2}{3}B}\right\}.$$

For a function g on Z, denote $E(g) = \int_Z g(z)d\rho$.

Lemma 3.2. *Let $\mathcal{G}$ be a set of functions on Z such that for some $c_\rho \geq 0$, $|g - E(g)| \leq B$ almost everywhere and $E(g^2) \leq c_\rho E(g)$ for each $g \in \mathcal{G}$. Then for every $\varepsilon > 0$ and $0 < \alpha \leq 1$,*

$$\mathrm{Prob}_{\mathbf{z}\in Z^m}\left\{\sup_{g\in\mathcal{G}} \frac{E(g) - \frac{1}{m}\sum_{i=1}^m g(z_i)}{\sqrt{E(g)+\varepsilon}} \geq 4\alpha\sqrt{\varepsilon}\right\} \leq \mathcal{N}(\mathcal{G}, \alpha\varepsilon)\exp\left\{-\frac{\alpha^2 m\varepsilon}{2c_\rho + \frac{2}{3}B}\right\}.$$

We apply Lemma 3.2 to a set of functions $\mathcal{F}_R$ with $R > 0$, where

$$\mathcal{F}_R = \left\{\big(f(x)-y\big)^2 - \big(f_\rho(x)-y\big)^2 : f \in B_R\right\}. \tag{3.1}$$

Proposition 3.1. *For all $\varepsilon > 0$ and $R \geq M$,*

$$\begin{aligned}\mathrm{Prob}_{\mathbf{z}\in Z^m}&\left\{\sup_{f\in B_R} \frac{\mathcal{E}(f) - \mathcal{E}(f_\rho) - \big(\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}_{\mathbf{z}}(f_\rho)\big)}{\sqrt{\mathcal{E}(f) - \mathcal{E}(f_\rho) + \varepsilon}} \leq \sqrt{\varepsilon}\right\} \\ &\geq 1 - \mathcal{N}\left(\frac{3\varepsilon}{4(\kappa+3)^2R^2}\right)\exp\left\{-\frac{3m\varepsilon}{160(\kappa+3)^2R^2}\right\}.\end{aligned}$$

Proof. Consider the set $\mathcal{F}_R$. Each function $g \in \mathcal{F}_R$ has the form $g(z) = \big(f(x)-y\big)^2 - \big(f_\rho(x)-y\big)^2$ with $f \in B_R$. Hence $E(g) = \mathcal{E}(f) - \mathcal{E}(f_\rho) \geq 0$, $\frac{1}{m}\sum_{i=1}^m g(z_i) = \mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}_{\mathbf{z}}(f_\rho)$, and

$$g(z) = (f(x) - f_\rho(x))\big\{\big(f(x)-y\big) + \big(f_\rho(x)-y\big)\big\}.$$

Since $\|f\|_\infty \leq \kappa\|f\|_K \leq \kappa R$ and $|f_\rho(x)| \leq M$ almost everywhere, we find that

$$|g(z)| \leq \big(\kappa R + M\big)\big(\kappa R + 3M\big) \leq c_R := (\kappa R + 3M)^2.$$

So we have $|g(z) - E(g)| \leq B := 2c_R$ almost everywhere.

Also,

$$E(g^2) = E\Big[(f(x)-f_\rho(x))^2\big\{\big(f(x)-y\big)+\big(f_\rho(x)-y\big)\big\}^2\Big] \le \big(\kappa R+3M\big)^2\|f-f_\rho\|_\rho^2.$$

Thus $E(g^2) \le c_R\|f-f_\rho\|_\rho^2 = c_R E(g)$ for each $g \in \mathcal{F}_R$.

Applying Lemma 3.2 with $\alpha = \frac{1}{4}$ to the function set $\mathcal{F}_R$, we deduce that

$$\sup_{f\in B_R} \frac{\mathcal{E}(f)-\mathcal{E}(f_\rho)-\big(\mathcal{E}_{\mathbf{z}}(f)-\mathcal{E}_{\mathbf{z}}(f_\rho)\big)}{\sqrt{\mathcal{E}(f)-\mathcal{E}(f_\rho)+\varepsilon}} = \sup_{g\in\mathcal{F}_R} \frac{E(g)-\frac{1}{m}\sum_{i=1}^m g(z_i)}{\sqrt{E(g)+\varepsilon}} \le \sqrt{\varepsilon}$$

with confidence at least

$$1-\mathcal{N}\left(\mathcal{F}_R,\varepsilon/4\right)\exp\left\{-\frac{m\varepsilon/16}{2c_R+\frac{2}{3}B}\right\} \ge 1-\mathcal{N}\left(\mathcal{F}_R,\varepsilon/4\right)\exp\left\{-\frac{3m\varepsilon}{160(\kappa+3)^2R^2}\right\}.$$

Here we have used the expressions for c_R, $B = 2c_R$ and the restriction $R \ge M$.

What is left is to bound the covering number $\mathcal{N}\left(\mathcal{F}_R,\varepsilon/4\right)$. To do so, note that

$$\big|\big(f_1(x)-y\big)^2-\big(f_2(x)-y\big)^2\big| = \|f_1-f_2\|_\infty\big|\big(f_1(x)-y\big)+\big(f_2(x)-y\big)\big|.$$

But $|y| \le M$ almost surely and $\|f\|_\infty \le \kappa R$ for each $f \in B_R$. Therefore,

$$\big|\big(f_1(x)-y\big)^2-\big(f_2(x)-y\big)^2\big| \le 2\big(M+\kappa R\big)\|f_1-f_2\|_\infty, \qquad \forall f_1,f_2\in B_R.$$

Since an $\frac{\eta}{2(MR+\kappa R^2)}$-covering of B_1 yields an $\frac{\eta}{2(M+\kappa R)}$-covering of B_R, and vice versa, we see that for any $\eta>0$, an $\frac{\eta}{2(MR+\kappa R^2)}$-covering of B_1 provides an η-covering of $\mathcal{F}_R$. Thus

$$\mathcal{N}(\mathcal{F}_R,\eta) \le \mathcal{N}\left(\frac{\eta}{2(MR+\kappa R^2)}\right), \qquad \forall \eta>0.$$

But $R \ge M$ and $8(1+\kappa) \le \frac{4}{3}(\kappa+3)^2$. So our desired estimate follows. □

§4. Error Bounds in a Weak Form

In this section we derive some weak error bounds in order to illustrate the idea by a simple procedure.

Proposition 4.1. *Suppose the kernel K satisfies (1.8), and $\mathcal{D}(\lambda) \le C_1\lambda^\beta$ for some $0 < \beta \le 1$ and $C_1 > 0$. For any $0 < \delta < 1$, with confidence $1-\delta$, there holds*

$$\|f_{\mathbf{z}} - f_\rho\|_\rho^2 \le \widetilde{C}\log\big(2/\delta\big)\left(\frac{1}{m}\right)^{\frac{\beta}{(1+\beta)(1+s)}}, \qquad \text{by taking } \lambda = \lambda(m) = m^{-\frac{1}{(1+\beta)(1+s)}}.$$

The proof of Proposition 4.1 follows from Corollary 4.1 and Proposition 4.3 below. Note the power $\frac{\beta}{(1+\beta)(1+s)}$ for the learning rate is less than $1/2$. Thus the estimate is weak. Strong error bounds will be given in the next section by more complicated arguments.

For the sample error estimates, we shall require the confidence $\mathcal{N}(\eta)\exp\big\{-\frac{m\eta}{40}\big\}$ to be δ. So we define the following quantity to realize this confidence.

Definition 4. *Let $g = g_{K,m} : \mathbb{R}_+ \to \mathbb{R}$ be the function given by*

$$g(\eta) = \log\mathcal{N}(\eta) - \frac{m\eta}{40}.$$

For $0 < \delta < 1$, we denote $v^*(m,\delta)$ as the unique solution to the equation

$$g(\eta) = \log\delta. \tag{4.1}$$

The function g is strictly decreasing in $(0,+\infty)$ with $g(0) = +\infty$ and $g(+\infty) = -\infty$. Also, $g(\kappa) = -m\kappa/40$. Therefore, the equation (4.1) has a solution for any $0 < \delta < 1$. Moreover,

$$\lim_{m\to\infty} v^*(m,\delta) = 0.$$

More quantitative decay estimates for $v^*(m,\delta)$ will be given at the end of this section.

Now we can derive the error bounds. For $R > 0$, denote

$$\mathcal{W}(R) = \{\mathbf{z} \in Z^m : \|f_{\mathbf{z}}\|_K \le R\}. \tag{4.2}$$

Proposition 4.2. *For all $0 < \delta < 1$ and $R \ge M$, there is a set $V_R \subseteq Z^m$ with $\rho(V_R) \le \delta$ such that for all $\mathbf{z} \in \mathcal{W}(R)\setminus V_R$, the regularized error $\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) + \lambda\|f_{\mathbf{z}}\|_K^2$ is bounded by*

$$3(\kappa+3)^2R^2v^*(m,\delta/2) + \widetilde{\mathcal{D}}(\lambda)\left(4 + \frac{8\kappa^2\log\big(2/\delta\big)}{m\lambda}\right) + \frac{72M^2\log\big(2/\delta\big)}{m}.$$

Proof. Note that $\big(\sqrt{\mathcal{E}(f) - \mathcal{E}(f_\rho) + \varepsilon}\big)\big(\sqrt{\varepsilon}\big) \le \frac{1}{2}\big(\mathcal{E}(f) - \mathcal{E}(f_\rho)\big) + \varepsilon$. Our statement follows directly via Propositions 1.2, 2.1, and 3.1 after replacing δ by $\delta/2$. □

Finally we need an R satisfying $\mathcal{W}(R) = Z^m$.

Lemma 4.1. *For all $\lambda > 0$ and almost all $\mathbf{z} \in Z^m$,*

$$\|f_{\mathbf{z}}\|_K \le \frac{M}{\sqrt{\lambda}}.$$

Proof. The definition of $f_{\mathbf{z}}$ tells us that for $f = 0$,

$$\lambda\|f_{\mathbf{z}}\|_K^2 \le \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \lambda\|f_{\mathbf{z}}\|_K^2 \le \mathcal{E}_{\mathbf{z}}(0) + 0 = \frac{1}{m}\sum_{i=1}^{m}(y_i - 0)^2 \le M^2$$

the last almost surely. Therefore, $\|f_{\mathbf{z}}\|_K \le M/\sqrt{\lambda}$ for almost all $\mathbf{z} \in Z^m$. □

Lemma 4.1 says that $\mathcal{W}(M/\sqrt{\lambda}) = Z^m$. Take $R := M/\sqrt{\lambda} \ge M$ when $0 < \lambda \le 1$. Then the following error bound follows from Proposition 4.2.

Corollary 4.1. *Let $0 < \lambda \le 1, \widetilde{f}_\lambda \in \mathcal{H}_K$ and $f_{\mathbf{z}}$ be defined by (1.4). Then for any $0 < \delta < 1$, with confidence $1 - \delta$, we have*

$$\|f_{\mathbf{z}} - f_\rho\|_\rho^2 \le 3(\kappa+3)^2M^2\frac{v^*(m,\delta/2)}{\lambda} + \widetilde{\mathcal{D}}(\lambda)\left(4 + \frac{8\kappa^2\log\big(2/\delta\big)}{m\lambda}\right) + \frac{72M^2\log\big(2/\delta\big)}{m}.$$

Learning rates in weak forms can be obtained from Corollary 4.1 and quantitative decay estimates for $v^*(m,\delta)$. The following estimate is essentially presented in [9]. For completeness, we give a proof.

Proposition 4.3. *If the kernel K satisfies (1.8), then*

$$v^*(m,\delta) \le \max\left\{\frac{80\log(1/\delta)}{m}, \big(80C_0/m\big)^{1/(1+s)}\right\}.$$

Proof. Observe that $g(\eta) \le h(\eta) := C_0(1/\eta)^s - \frac{m\eta}{40}$. Then

$$\log\delta = g(v^*(m,\delta)) \le h(v^*(m,\delta)).$$

Since h is also strictly decreasing, we know that $v^*(m,\delta) \le \Delta$ where Δ is the unique positive solution of the equation $h(t) = \log\delta$. This new equation can be expressed as

$$t^{1+s} - \frac{40\log(1/\delta)}{m}t^s - \frac{40C_0}{m} = 0.$$

Then Lemma 7 from [9] yields $\Delta \le \max\big\{\frac{80\log(1/\delta)}{m}, \big(80C_0/m\big)^{1/(1+s)}\big\}$. This verifies the bound for $v^*(m,\delta)$. □

§5. Strong Estimates by Iteration

In this section we improve the error estimate stated in Corollary 4.1. Our main result here can be stated as the following theorem which will be proved after two lemmas.

Theorem 2. *Let $f_{\mathbf{z}}$ be defined by (1.4). Assume (1.8). Set $C_2 = (2\kappa+6)\big(80C_0\big)^{1/(2+2s)} + 1$. Take $\lambda = m^{\epsilon-\frac{1}{1+s}}$ with $0 < \epsilon \le \frac{1}{1+s}$. For any $0 < \delta < 1$ and $m \ge m_{\delta,\epsilon}$, with confidence $1-\delta$ there holds*

$$\|f_{\mathbf{z}} - f_\rho\|_\rho^2 \le C_\epsilon \log\big(\frac{2}{\delta}\big) \left(\mathcal{D}(\lambda) + \frac{2}{m^{1/(1+s)}}\right)$$

where $m_{\delta,\epsilon}$ and C_ϵ are constants given by

$$m_{\delta,\epsilon} = \max\left\{ \frac{80}{C_0^{1/s}} \left(\log\frac{2}{\delta} + \log\frac{1+\epsilon(1+s)}{\epsilon(1+s)}\right)^{1+1/s}, \left(\frac{1}{2C_2}\right)^{2/\epsilon} \right\},$$

$$C_\epsilon = \log\frac{1+\epsilon(1+s)}{\epsilon(1+s)} \Bigg\{ 12(\kappa+3)^2 \left(6M + MC_2^{\frac{1}{\epsilon(1+s)}} + 2\kappa + 2\right)^2 (80C_0)^{\frac{1}{1+s}} + 4 + 8\kappa^2 + 72M^2 \Bigg\}.$$

The method in the previous section was rough because we used the bound $\|f_{\mathbf{z}}\|_K \le M/\sqrt{\lambda}$ shown in Lemma 4.1. This is much worse than the bound for f_λ derived from the proof of Proposition 2.1, namely, $\|f_\lambda\|_K \le \sqrt{\mathcal{D}(\lambda)}/\sqrt{\lambda}$. Yet, we expect $f_{\mathbf{z}}$ to be a good approximation of f_λ. In particular, one would expect $\|f_{\mathbf{z}}\|_K$ to be bounded as well by, essentially, $\sqrt{\mathcal{D}(\lambda)}/\sqrt{\lambda}$. We shall prove that this is the case with high probability by applying Proposition 4.2 iteratively. As a consequence, we will obtain strong error estimates. The iteration technique was introduced in [16] and improved in [24] for the purpose of support vector machine classification algorithms. Our iteration technique here refines and is different from the previous ones. In particular, the method in [16] requires a polynomial decay of the regularization error $\mathcal{D}(\lambda) = O(\lambda^\beta)$ with some $0 < \beta \le 1$. Similar ideas of norm reduction also appear in [13] for the purpose of bounding the risk of function learning.

Recall the set $\mathcal{W}(R)$ defined by (4.2). Proposition 4.2 immediately yields the following relation, since $\lambda\|f_{\mathbf{z}}\|_K^2$ is bounded by the regularized error $\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) + \lambda\|f_{\mathbf{z}}\|_K^2$.

Lemma 5.1. *For any $0 < \delta < 1$ and $R \geq M$, there is a set $V_R \subseteq Z^m$ with $\rho(V_R) \leq \delta$ such that*

$$\mathcal{W}(R) \subseteq \mathcal{W}(a_m R + b_m) \cup V_R,$$

where a_m, b_m are independent of R and given by $a_m := (2\kappa + 6)\sqrt{v^*(m, \delta/2)/\lambda}$,

$$b_m := \left(\frac{2\kappa\sqrt{2\log\big(2/\delta\big)}}{\sqrt{m\lambda}} + 2 \right) \sqrt{\widetilde{\mathcal{D}}(\lambda)/\lambda} + \frac{6M\sqrt{2\log\big(2/\delta\big)}}{\sqrt{m\lambda}}.$$

Let us describe our iteration procedure briefly. According to Proposition 4.3, when (1.8) is valid, $v^*(m, \delta/2) \leq \big(80C_0/m\big)^{1/(1+s)}$ for sufficiently large m (depending on δ). Thus if we choose $\lambda = m^{\epsilon - \frac{1}{1+s}}$ in Lemma 5.1, $a_m = O(m^{-\epsilon/2})$ and Lemma 5.1 ensures a reduction of the norm $\|f_{\mathbf{z}}\|_K$: the set $\mathcal{W}(R)$ coincides with $\mathcal{W}(R')$ with a smaller radius $R' = O\left(m^{-\epsilon/2} R + \sqrt{\widetilde{\mathcal{D}}(\lambda)/\lambda} \right)$, up to a small set V_R of measure at most δ. Therefore, starting with $\mathcal{W}(R^{(0)}) = Z^m$ and $R^{(0)} = M/\sqrt{\lambda}$ proved in Lemma 4.1, we iterate the above reduction procedure J times and can find that up to a set of measure at most $J\delta$, the sets $Z^m = \mathcal{W}(R^{(0)})$ and $\mathcal{W}(R^{(J)})$ coincide. The radius $R^{(J)}$ can be bounded by $O\left(m^{-J\epsilon/2} R^{(0)} + \sqrt{\widetilde{\mathcal{D}}(\lambda)/\lambda} \right) = O\left(m^{-J\epsilon/2 - \epsilon/2 + 1/(2+2s)} + \sqrt{\widetilde{\mathcal{D}}(\lambda)/\lambda} \right)$. Hence the choice $J \geq \frac{1}{\epsilon(1+s)} - 1$ ensures the essential bound $R^{(J)}$ of $\|f_{\mathbf{z}}\|_K$ to be $\sqrt{\widetilde{\mathcal{D}}(\lambda)/\lambda}$ with probability at least $1 - J\delta$.

Lemma 5.2. *Assume (1.8). Take $\lambda = m^{\epsilon - \frac{1}{1+s}}$ with $0 < \epsilon \leq \frac{1}{1+s}$. For any $0 < \delta < 1$ and $m \geq m'_\delta$, with confidence $1 - \frac{\delta}{\epsilon(1+s)}$ there holds*

$$\|f_{\mathbf{z}}\|_K \leq \left(6M + MC_2^{\frac{1}{\epsilon(1+s)}} + 2\kappa + 2 \right) \sqrt{2\log\big(2/\delta\big)} \left(\sqrt{\widetilde{\mathcal{D}}(\lambda)/\lambda} + 1 \right).$$

Here $m'_\delta := \max\left\{ \frac{80}{C_0^{1/s}} \big(\log(2/\delta)\big)^{1+1/s}, \big(1/(2C_2)\big)^{2/\epsilon} \right\}$.

Proof. By Proposition 4.3, when $m \geq \frac{80}{C_0^{1/s}} \big(\log(2/\delta)\big)^{1+1/s}$, there holds

$$v^*(m, \delta/2) \leq \big(80C_0/m\big)^{1/(1+s)}. \tag{5.1}$$

It follows that

$$a_m \leq (2\kappa + 6)\big(80C_0\big)^{1/(2+2s)} m^{-\epsilon/2} \leq C_2\, m^{-\epsilon/2}.$$

If m satisfies a second restriction $m \geq \big(1/(2C_2)\big)^{2/\epsilon}$, we have $a_m \leq 1/2$. Thus,

$$a_m \leq C_2\ m^{-\epsilon/2} \leq 1/2, \qquad \forall m \geq m'_\delta. \tag{5.2}$$

Define a sequence $\{R^{(j)}\}_{j\in\mathbb{N}}$ by $R^{(0)} = M/\sqrt{\lambda}$ and, for $j \geq 1$,

$$R^{(j)} = a_m R^{(j-1)} + b_m.$$

Then Lemma 4.1 proves $\mathcal{W}(R^{(0)}) = Z^m$, and Lemma 5.1 asserts that for each $j \geq 1$, $\mathcal{W}(R^{(j-1)}) \subseteq \mathcal{W}(R^{(j)}) \cup V_{R^{(j-1)}}$ with $\rho(V_{R^{(j-1)}}) \leq \delta$. Apply this inclusion for $j = 1, 2, \ldots, J$, with J satisfying $\frac{1}{\epsilon(1+s)} - 1 \leq J \leq \frac{1}{\epsilon(1+s)}$. We see that

$$Z^m = \mathcal{W}(R^{(0)}) \subseteq \mathcal{W}(R^{(1)}) \cup V_{R^{(0)}} \subseteq \cdots \subseteq \mathcal{W}(R^{(J)}) \cup \left(\bigcup_{j=0}^{J-1} V_{R^{(j)}}\right).$$

It follows that the measure of the set $\mathcal{W}(R^{(J)})$ is at least $1 - J\delta \geq 1 - \frac{\delta}{\epsilon(1+s)}$. By the definition of the sequence, we have

$$R^{(J)} = a_m^J R^{(0)} + b_m \sum_{j=0}^{J-1} a_m^j.$$

By (5.2), $a_m \leq 1/2$, hence $\sum_{j=0}^{J-1} a_m^j \leq 1$. The bound $a_m \leq C_2 m^{-\epsilon/2}$ in (5.2) and $R^{(0)} = M/\sqrt{\lambda} = M m^{\frac{1}{2+2s} - \frac{\epsilon}{2}}$ yield

$$a_m^J R^{(0)} \leq C_2^J m^{-J\epsilon/2} M m^{\frac{1}{2+2s} - \frac{\epsilon}{2}} = C_2^J M m^{\frac{1}{2+2s} - \frac{\epsilon}{2} - \frac{J\epsilon}{2}}.$$

But $J \geq \frac{1}{\epsilon(1+s)} - 1$ which implies $\frac{1}{2+2s} - \frac{\epsilon}{2} - \frac{J\epsilon}{2} \leq 0$. Hence $a_m^J R^{(0)} \leq C_2^J M$.

Since $m\lambda \geq 1$, b_m can be bounded as

$$b_m \leq \sqrt{2\log\big(2/\delta\big)}\left((2\kappa+2)\sqrt{\widetilde{\mathcal{D}}(\lambda)/\lambda} + 6M\right).$$

Therefore

$$R^{(J)} \leq \left(6M + MC_2^J + 2\kappa + 2\right)\sqrt{2\log\big(2/\delta\big)}\big(\sqrt{\widetilde{\mathcal{D}}(\lambda)/\lambda} + 1\big).$$

This proves our statement. □

We are in a position to prove our main results.

Proof of Theorem 2. Choose $\widetilde{f}_\lambda = f_\lambda$. Then $\widetilde{\mathcal{D}}(\lambda) = \mathcal{D}(\lambda)$. For $0 < \delta < 1$, let $\tilde{\delta} := \frac{\epsilon(1+s)}{1+\epsilon(1+s)}\delta \in (0,1)$ and $m_{\delta,\epsilon} = m'_{\tilde{\delta}}$ be as in Lemma 5.2. Take

$$R = \left(6M + MC_2^{\frac{1}{\epsilon(1+s)}} + 2\kappa + 2\right)\sqrt{2\log\big(2/\tilde{\delta}\big)}\left(\sqrt{\mathcal{D}(\lambda)/\lambda} + 1\right).$$

Let $m \geq m_{\delta,\epsilon}$. Lemma 5.2 tells us that the measure of the set $\mathcal{W}(R)$ is at least $1 - \frac{\tilde{\delta}}{\epsilon(1+s)}$.

Applying Proposition 4.2 to the above R, we know that for each $\mathbf{z} \in \mathcal{W}(R) \setminus V_R$,

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) \leq 12(\kappa+3)^2\left(6M + MC_2^{\frac{1}{\epsilon(1+s)}} + 2\kappa + 2\right)^2 \log\big(\frac{2}{\tilde{\delta}}\big)\left(\frac{\mathcal{D}(\lambda)}{\lambda} + 1\right) v^*(m, \frac{\tilde{\delta}}{2})$$
$$+ \mathcal{D}(\lambda)\left(4 + \frac{8\kappa^2\log\big(2/\tilde{\delta}\big)}{m\lambda}\right) + \frac{72M^2\log\big(2/\tilde{\delta}\big)}{m}.$$

Since the measure of V_R is at most $\tilde{\delta}$, we know that the above error bound holds for $\mathbf{z} \in \mathcal{W}(R) \setminus V_R$ which has measure at least $1 - \frac{\tilde{\delta}}{\epsilon(1+s)} - \tilde{\delta} = 1 - \delta$. Putting the bound (5.1) into the above error estimate and noting that $\log\big(2/\tilde{\delta}\big) \leq \log\big(2/\delta\big) + \log\frac{1+\epsilon(1+s)}{\epsilon(1+s)}$, we see our conclusion. □

Now we can prove Theorem 1 stated in the introduction.

Proof of Theorem 1. Take $\epsilon = \frac{1}{1+s} - \zeta$. Then $\lambda = m^{\epsilon - \frac{1}{1+s}}$, and the conclusion of Theorem 2 holds.

When f_ρ is not identically zero, then $\|f_\rho\|_\rho > 0$. It follows from (1.3) that

$$\mathcal{D}(\lambda) \geq \inf_{f \in \mathcal{H}_K}\left\{\|f - f_\rho\|_\rho^2 + \frac{\lambda}{\kappa^2}\|f\|_\rho^2\right\} \geq \frac{\lambda}{4\kappa^2}\|f_\rho\|_\rho^2.$$

Therefore,

$$\frac{2}{m^{1/(1+s)}} \leq 2\lambda \leq \frac{8\kappa^2}{\|f\|_\rho^2}\mathcal{D}(\lambda).$$

Let $m_{\delta,\zeta} := m_{\delta,\epsilon}$ be given in Theorem 2. Note that $\|f_\rho\|_\rho \leq M$. The error bound in Theorem 2 tells us that for $m \geq m_{\delta,\zeta}$ there holds

$$\|f_{\mathbf{z}} - f_\rho\|_\rho^2 \leq C_\epsilon \log\big(\frac{2}{\delta}\big)\big(1 + \frac{8\kappa^2}{\|f\|_\rho^2}\big)\mathcal{D}(\lambda) \leq \frac{C_\epsilon}{\|f\|_\rho^2}\big(M^2 + 8\kappa^2\big)\log\big(\frac{2}{\delta}\big)\mathcal{D}(\lambda)$$

with confidence $1 - \delta$. This proves Theorem 1 with $\widetilde{C} = \big(M^2 + 8\kappa^2\big)C_\epsilon$. □

To show that the constant term in (1.9) depends on $\|f_\rho\|_\rho^2$, we consider an example with the simplest Mercer kernel ($K \equiv 1$).

Example. *Let $K \equiv 1$ and $f_{\mathbf{z}}$ be defined by (1.4). If $f_\rho(x) \equiv \mu_\rho$ for some $\mu_\rho \in (-M, M)$, then $\mathcal{D}(\lambda) = \frac{\lambda}{1+\lambda}\mu_\rho^2$ and*

$$E_{\mathbf{z}\in Z^m}\big(\|f_{\mathbf{z}} - f_\rho\|_\rho^2\big) = \frac{\mathcal{E}(f_\rho)}{(1+\lambda)^2 m} + \frac{\lambda^2}{(1+\lambda)^2}\mu_\rho^2.$$

Proof. The space $\mathcal{H}_K$ consists of constant functions $f \equiv c$ with $\|f\|_K = |c|$. Then $\mathcal{E}_{\mathbf{z}}(f) + \lambda\|f\|_K^2 = \frac{1}{m}\sum_{i=1}^m y_i^2 - \frac{2c}{m}\sum_{i=1}^m y_i + (1+\lambda)c^2$ and

$$f_{\mathbf{z}} \equiv \frac{1}{(1+\lambda)m}\sum_{i=1}^m y_i.$$

Since $f_\rho(x) \equiv \mu_\rho$, we have $E(y) = \mu_\rho$ and $E\{(y-\mu_\rho)^2\} = \mathcal{E}(f_\rho)$. It follows that

$$E_{\mathbf{z}\in Z^m}\big(\|f_{\mathbf{z}} - f_\rho\|_\rho^2\big) = E_{\mathbf{z}\in Z^m}\left\{\Big(\frac{1}{(1+\lambda)m}\sum_{i=1}^m (y_i - \mu_\rho) - \frac{\lambda}{1+\lambda}\mu_\rho\Big)^2\right\}$$

verifying the desired expected value.

For any $\lambda > 0$, we have $\mathcal{D}(\lambda) = \inf_{c\in\mathbb{R}}\big\{(c-\mu_\rho)^2 + \lambda c^2\big\} = \frac{\lambda}{1+\lambda}\mu_\rho^2$. □

In the above example, (1.8) holds for any $s > 0$. If we take $\lambda = m^{-\zeta}$ with $0 < \zeta < 1$, we have

$$\frac{E_{\mathbf{z}\in Z^m}\big(\|f_{\mathbf{z}} - f_\rho\|_\rho^2\big)}{\mathcal{D}(\lambda)} \ge \frac{\mathcal{E}(f_\rho)m^{\zeta-1}}{2\|f_\rho\|_\rho^2}, \qquad \forall m \in \mathbb{N}.$$

In particular, when $\mathcal{E}(f_\rho) > 0$ and $m_{\delta,\zeta}$, $\widetilde{C}$ are constants depending only on C_0, s, ζ, κ, M and δ, by letting $\|f_\rho\|_\rho^2 = \mu_\rho^2 \to 0$ we see that the bound (1.9) without the denominator $\|f_\rho\|_\rho^2$ in the constant term cannot be true for all $m \ge m_{\delta,\zeta}$.

§6. Deriving Learning Rates

Our main result, Theorem 1, on the error analysis yields learning rates.

Define an integral operator $L_K : L^2_{\rho_X} \to L^2_{\rho_X}$ by

$$L_K(f)(x) = \int_X K(x,y)f(y)d\rho_X(y), \qquad x \in X.$$

Its range is in $\mathcal{H}_K$. The operator L_K can also be defined as a self-adjoint operator on $\mathcal{H}_K$ or on $L^2_{\rho_X}$. We shall use the same notion L_K for these operators defined on different domains.

It was shown in [8, Theorem 3] and [20, (7.11)] that when $L_K^{-r} f_\rho \in L^2_{\rho_X}$ for some $0 < r \le 1/2$, there holds $\mathcal{D}(\lambda) \le \lambda^{2r}\|L_K^{-r} f_\rho\|_\rho^2$. This in connection with Theorem 1 verifies the following.

Corollary 6.1. *Let $f_{\mathbf{z}}$ be defined by (1.4). Assume (1.8). If $L_K^{-r} f_\rho \in L^2_{\rho_X}$ for some $0 < r \le 1/2$, then for any $0 < \zeta < \frac{1}{1+s}$, $0 < \delta < 1$ and $m \ge m_{\delta,\epsilon}$, with confidence $1-\delta$*

$$\|f_{\mathbf{z}} - f_\rho\|_\rho^2 \le \widetilde{C} \log\left(\frac{2}{\delta}\right) m^{-2r\zeta}$$

where $m_{\delta,\epsilon}$ and $\widetilde{C}$ are constants independent of m.

In particular, when K is C^∞ on $X \subset \mathbb{R}^n$, we know from [30] that (1.8) holds for any $s > 0$. As a consequence of Corollary 6.1, we see the following learning rates.

Corollary 6.2. *Let $f_{\mathbf{z}}$ be defined by (1.4). If K is C^∞ on $X \subset \mathbb{R}^n$ and $L_K^{-r} f_\rho \in L^2_{\rho_X}$ for some $0 < r \le 1/2$, then for any $\epsilon > 0, 0 < \delta < 1$, with confidence $1-\delta$ there holds*

$$\|f_{\mathbf{z}} - f_\rho\|_\rho^2 \le \widetilde{C} \log(\frac{2}{\delta}) \left(\frac{1}{m}\right)^{2r-\epsilon}$$

for $m \ge m_{\delta,\epsilon}$ and $\lambda = \lambda(m) = m^{2\epsilon-1}$, where $m_{\delta,\epsilon}$ and $\widetilde{C}$ are constants independent of m.

When $r = 1/2$, $f_\rho \in \mathcal{H}_K$, the learning rates in Corollary 6.2 is arbitrarily close to 1, which was stated in Proposition 1.1.

Let us compare our learning rates with the existing results.

In [6, 28], a leave-one-out technique was used to derive the expected value of learning schemes. For the scheme (1.4), the result in [28] can be expressed as

$$E_{\mathbf{z}\in Z^m}\big(\mathcal{E}(f_{\mathbf{z}})\big) \le \left(1 + \frac{2\kappa^2}{m\lambda}\right)^2 \inf_{f\in\mathcal{H}_K}\left\{\mathcal{E}(f) + \frac{\lambda}{2}\|f\|_K^2\right\}.$$

In terms of the regularization error, it can be restated as

$$E_{\mathbf{z}\in Z^m}\big(\big\|f_{\mathbf{z}} - f_\rho\big\|_\rho^2\big) \le \mathcal{D}\big(\lambda/2\big) + \Big(\mathcal{E}(f_\rho) + \mathcal{D}\big(\lambda/2\big)\Big)\left\{\frac{4\kappa^2}{m\lambda} + \left(\frac{2\kappa^2}{m\lambda}\right)^2\right\}.$$

Choosing $\lambda = 1/\sqrt{m}$, the derived learning rate is $\big(\frac{1}{m}\big)^{\frac{1}{2}}$ in expectation when $f_\rho \in \mathcal{H}_K$ and $\mathcal{E}(f_\rho) > 0$. By the Markov inequality, $\big\|f_{\mathbf{z}} - f_\rho\big\|_\rho^2 \le \frac{C}{\delta}\big(\frac{1}{m}\big)^{\frac{1}{2}}$ with confidence $1-\delta$.

In [11], a functional analysis approach was employed for the error analysis of the scheme (1.4). The main result asserts that for any $0 < \delta < 1$, with confidence $1-\delta$,

$$\big|\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\lambda)\big| \le \frac{M\kappa^2}{\sqrt{m}\lambda}\left(1 + \frac{\kappa}{\sqrt{\lambda}}\right)\left(1 + \sqrt{2\log\big(2/\delta\big)}\right).$$

The corresponding learning rate [11, Corollary 1] is the following: when f_ρ lies in the range of L_K, i.e., $L_K^{-1} f_\rho \in L^2_{\rho_X}$, for any $0 < \delta < 1$, with confidence $1 - \delta$, there holds

$$\big\|f_{\mathbf{z}} - f_\rho\big\|_\rho^2 \le C\big(\frac{\log\big(2/\delta\big)}{m}\big)^{\frac{2}{5}}, \qquad \text{if} \qquad \lambda = \big(\frac{\log\big(2/\delta\big)}{m}\big)^{\frac{1}{5}}.$$

Thus the confidence is improved from $1/\delta$ to $\log(2/\delta)$, while the rate is weakened to $\big(\frac{1}{m}\big)^{\frac{2}{5}}$.

In [20], a modified McDiarmid inequality was used to improve the kernel independent error bounds. If f_ρ is in the range of L_K, then for any $0 < \delta < 1$, with confidence $1 - \delta$ there hold

$$\|f_{\mathbf{z}} - f_\rho\|_K^2 \le \widetilde{C}\bigg(\frac{\big(\log\big(4/\delta\big)\big)^2}{m}\bigg)^{\frac{1}{3}} \qquad \text{by taking} \qquad \lambda = \bigg(\frac{\big(\log\big(4/\delta\big)\big)^2}{m}\bigg)^{\frac{1}{3}} \tag{6.1}$$

and

$$\|f_{\mathbf{z}} - f_\rho\|_\rho^2 \le \widetilde{C}\frac{\log\big(4/\delta\big)}{\sqrt{m}} \qquad \text{by taking} \qquad \lambda = \bigg(\frac{\big(\log\big(4/\delta\big)\big)^2}{m}\bigg)^{\frac{1}{4}}, \tag{6.2}$$

Thus the confidence for the learning rate $m^{-\frac{1}{2}}$ is improved. Moreover, the error in the space $\mathcal{H}_K$ can be estimated.

All the above results are kernel independent error bounds. When $f_\rho \in \mathcal{H}_K$ and $s < 1$, the learning rate given by Corollary 6.1 with $r = 1/2$ is better than existing results. In particular, this is the case [30] when the kernel K is C^p, with $p > 2n$, on $X \subset \mathbb{R}^n$.

Classical results [17, 14, 2, 8] on analysis of empirical risk minimization (ERM) schemes give error bounds between the empirical target function (over a bounded hypothesis space of functions) and the regression function. In particular, learning rates of type $m^{-\zeta}$ with ζ arbitrarily close to 1 can be achieved by ERM schemes. See e.g. [14, 2, 8]. However, the ERM setting is different from the one on Tikhonov regularization. How to choose the regularization parameter $\lambda = \lambda(m)$, depending on the sample size m, is the essential difficulty for the regularization scheme, even when f_ρ lies in $\mathcal{H}_K$.

§7. Extensions: Projection, Empirical Covering, and Multi-kernel Learning

Our error analysis can be extended in different settings.

The first extension is to use the projection. By our assumption, $f_\rho(x) \in [-M, M]$. Thus, it's natural for us to restrict approximating functions onto those supported also on $[-M, M]$.

Definition 5. *The projection operator π_M is defined on the space of measurable functions $f : X \to \mathbb{R}$ as*

$$\pi_M(f)(x) = \begin{cases} M, & \text{if } f(x) > M, \\ -M, & \text{if } f(x) < -M, \\ f(x), & \text{if } -M \le f(x) \le M. \end{cases} \tag{7.1}$$

The idea of projections appeared in margin-based bound analysis, e.g. [5, 15]. We introduced the above form of the projection operator π_1 for the purpose of bounding misclassification and generalization errors in [7], and used it for error analysis of linear programming support vector machine classification algorithms [26].

Now we take $\pi_M(f_{\mathbf{z}})$ as our empirical target function. Then (1.11) still holds after replacing $f_{\mathbf{z}}$ by $\pi_M(f_{\mathbf{z}})$:

$$\begin{aligned} \mathcal{E}(\pi_M(f_{\mathbf{z}})) - \mathcal{E}(f_\rho) + \lambda \|f_{\mathbf{z}}\|_K^2 \le & \left\{ \mathcal{E}(\widetilde{f}_\lambda) - \mathcal{E}(f_\rho) + \lambda \|\widetilde{f}_\lambda\|_K^2 \right\} \\ & + \left\{ \mathcal{E}(\pi_M(f_{\mathbf{z}})) - \mathcal{E}_{\mathbf{z}}(\pi_M(f_{\mathbf{z}})) + \mathcal{E}_{\mathbf{z}}(\widetilde{f}_\lambda) - \mathcal{E}(\widetilde{f}_\lambda) \right\}. \end{aligned}$$

Then our analysis can be done in the same way. The projection helps us to get sharper bounds of the sample error: probability inequalities are applied to random variables involving functions $\pi_M(f_{\mathbf{z}})$ (bounded by M), not $f_{\mathbf{z}}$ (the corresponding bound increases to infinity as λ becomes small, as shown in Lemmas 4.1 or 5.2). We omit details for deriving learning rates.

The second extension is to use empirical covering numbers, not the uniform covering numbers.

Definition 6. *Let $\mathcal{F}$ be a class of functions on X and $\mathbf{x} = \{x_1, \cdots, x_m\} \subset X$. For $1 \le p \le \infty$, define*

$$d_{p,\mathbf{x}}(f, g) = \left\{ \frac{1}{m} \sum_{i=1}^m |f(x_i) - g(x_i)|^p \right\}^{1/p}.$$

For every $\varepsilon > 0$, the covering number of $\mathcal{F}$ associated to $d_{p,\mathbf{x}}$ is

$$\mathcal{N}_{p,\mathbf{x}}(\mathcal{F}, \varepsilon) = \min \left\{ \ell \in \mathbb{N} : \exists \{f_j\}_{j=1}^\ell \text{ such that } \mathcal{F} = \bigcup_{j=1}^\ell \{f \in \mathcal{F} : d_{p,\mathbf{x}}(f, f_j) \le \varepsilon\} \right\}.$$

The p-empirical covering number of $\mathcal{F}$ is then defined by

$$\mathcal{N}_p(\mathcal{F}, \varepsilon) = \sup_{m \in \mathbb{N}, \mathbf{x} \in X^m} \mathcal{N}_{p,\mathbf{x}}(\mathcal{F}, \varepsilon).$$

The mostly used cases in learning theory are $p = 1,\ 2$ and ∞. One can easily see that

$$\mathcal{N}_1(\mathcal{F}, \varepsilon) \le \mathcal{N}_2(\mathcal{F}, \varepsilon) \le \mathcal{N}_\infty(\mathcal{F}, \varepsilon) \le \mathcal{N}(\mathcal{F}, \varepsilon).$$

Hence one may expect to use empirical covering numbers to get better bounds for the sample error and hence sharper learning rates. In particular, when $p = 2$, one may use techniques from empirical process theory such as a chaining argument and Dudley's entropy integral and other probability inequalities such as Talagrand's inequality to have better estimates for the sample error (e.g. [24]). Moreover, the covering number of the reproducing kernel Hilbert space satisfies $\log \mathcal{N}_2(B_1, \varepsilon) \le c(1/\varepsilon)^s$ with $0 < s \le 2$ (see e.g. [21]). Actually, there holds $\int_0^1 \sqrt{\log \mathcal{N}_2(B_1, \varepsilon)} d\varepsilon < \infty$. When s tends to the extremal case 2, learning rates obtained by these techniques can be arbitrarily close to those deduced by kernel independent bounds based on the leave-one-out analysis.

On the other hand, Mercer kernels may have logarithmic complexity exponent $s \ge 1$: for some $C_3 > 0$, there holds

$$\log \mathcal{N}(\eta) \le C_3 \left(\log(1/\eta)\right)^s, \qquad \forall \eta > 0. \tag{7.2}$$

This is better than (1.8). In particular, for convolution type kernels $K(x, y) = k(x - y)$ generated by real even functions k having exponentially decaying Fourier transform, (7.2) holds. As an example, consider the Gaussian kernel $K(x, y) = \exp\{-|x - y|^2/\sigma^2\}$ with $\sigma > 0$. If $X \subset [0, 1]^n$ and $0 < \eta \le \exp\{90n^2/\sigma^2 - 11n - 3\}$, then (7.2) is valid [29] with $s = n + 1$. A lower bound [30] holds with $s = \frac{n}{2}$, which shows the upper bound is almost sharp.

When the kernel has logarithmic complexity exponent $s \ge 1$ with (7.2) satisfied, then

$$v^*(m, \delta) \le \left(40 \log(1/\delta) + 40C_3 \big(\log m\big)^s + 1\right)/m.$$

To see this, we use the same procedure as Proposition 4.3. Take $\tilde{h}(\eta) := C_3 \big(\log \frac{1}{\eta}\big)^s - \frac{m\eta}{40}$. Then $v^*(m, \delta) \le \Delta$ where Δ is any positive number satisfying $\tilde{h}(\Delta) \le \log \delta$. We can easily check that

$$\tilde{h}\bigg(\big(40 \log(1/\delta) + 40C_3 \big(\log m\big)^s + 1\big)/m\bigg) \le C_3 \big(\log m\big)^s - C_3 \big(\log m\big)^s - \log \frac{1}{\delta} = \log \delta.$$

Then the bound for $v^*(m,\delta)$ follows. The small capacity kernels (7.2) yield learning rates of the type $O((\log m)^s/m)$. This can be derived by a refined iteration, which is presented in this paper.

The third extension is the multi-kernel setting. In this case, a set of kernels are used instead of only one kernel. Let Σ be an index set and $\{K_\sigma\}_{\sigma\in\Sigma}$ a set of Mercer kernels. The multi-kernel regularized learning scheme is defined as

$$f_{\mathbf{z}} = f_{\mathbf{z},\lambda} := \arg\min_{\sigma\in\Sigma} \min_{f\in\mathcal{H}_{K_\sigma}} \Big\{\mathcal{E}_{\mathbf{z}}(f) + \lambda\|f\|^2_{K_\sigma}\Big\}.$$

Particular examples of kernel sets include Gaussians (see [27]) with flexible variances $\big\{K_\sigma\big\}_{0<\sigma<\infty}$, $K_\sigma(x,y) = \exp\big\{-\frac{|x-y|^2}{\sigma^2}\big\}$, and polynomial kernels (see [32]) with varying degrees $\{K_d\}_{d\in\mathbb{N}}$, $K_d(x,y) = (1+x\cdot y)^d$. The main advantage of this multi-kernel algorithm is to improve the regularization error by using varying hypothesis spaces (see [31, 16, 32]). The error analysis for this multi-kernel setting can be done in the same way if the covering number of $\bigcup_{\sigma\in\Sigma}\big\{f\in\mathcal{H}_{K_\sigma} : \|f\|_{K_\sigma}\le 1\big\}$ satisfies (1.8). But the index s may be large $s>2$, due to the multi-kernels. It's unknown whether the empirical covering number or some other data dependent capacity measurements [17, 1] can be used in this setting.

Acknowledgments

The work is supported by the Research Grants Council of Hong Kong [Project No. CityU 1087/02P]. The second author is on leave from Institute of Mathematics, Chinese Academy of Science, Beijing 100080, CHINA. The authors would like to thank the referees for their careful reading and helpful comments on the paper.

References

[1] S. Andonova, Generalization bounds and complexities based on sparsity and clustering for convex combinations of functions from random classes, J. Machine Learning Res. **6** (2005), 307–340.

[2] M. Anthony, and P. L. Bartlett, Neural Network Learning: Theoretical Foundations, Cambridge University Press, 1999.

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

[4] A. R. Barron, Complexity regularization with applications to artificial neural networks, in *Nonparametric Functional Estimation* (G. Roussa, ed.), Kluwer, Dortrecht, 1990, pp. 561–576.

[5] 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.

[6] O. Bousquet and A. Elisseeff, Stability and generalization, J. Mach. Learning Res. **2** (2002), 499–526.

[7] D. R. Chen, Q. Wu, Y. Ying, and D. X. Zhou, Support vector machine soft margin classifiers: error analysis, J. Machine Learning Research **5** (2004), 1143–1175.

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

[9] F. Cucker and S. Smale, Best choices for regularization parameters in learning theory: On the bias-variance problem, Found. Comput. Math. **2** (2002), 413–428.

[10] F. Cucker and D. X. Zhou, Learning Theory: an Approximation Theory Viewpoint, monograph manuscript in preparation for Cambridge University Press.

[11] E. De Vito, A. Caponnetto, and L. Rosasco, Model selection for regularized least-squares algorithm in learning theory, Found. Comput. Math. **5** (2005), 59–85. .

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

[13] V. Koltchinskii and D. Panchenko, Rademacher processes and bounding the risk of function learning, in "High Dimensional Probability II", E. Gine, D. M. Mason, J. A. Wellner (eds.), Birkhäuser, Boston, 2000, pp. 443 - 459.

[14] W. S. Lee, P. Bartlett, and R. Williamson, The importance of convexity in learning with least square loss, IEEE Trans. Inform. Theory **44** (1998), 1974–1980.

[15] G. Lugosi and N. Vayatis, On the Bayes-risk consistency of regularized boosting methods, Ann. Stat. **32** (2004), 30–55.

[16] C. Scovel and I. Steinwart, Fast rates for support vector machines, preprint, 2003.

[17] J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony, Structural risk minimization over data-dependent hierarchies, IEEE Trans. Inform. Theory **44** (1998), 1926–1940.

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

[19] S. Smale and D. X. Zhou, Shannon sampling and function reconstruction from point values, Bull. Amer. Math. Soc. **41** (2004), 279–305.

[20] S. Smale and D. X. Zhou, Shannon sampling II. Connections to learning theory, Appl. Comput. Harmonic Anal., to appear.

[21] A. W. van der Vaart and J. A. Wellner, Weak Convergence and Empirical Processes, Springer-Verlag, New York, 1996.

[22] V. Vapnik, Statistical Learning Theory, John Wiley & Sons, 1998.

[23] G. Wahba, Spline models for observational data, SIAM, 1990.

[24] Q. Wu, Y. Ying, and D. X. Zhou, Multi-kernel regularized classifiers, submitted, 2004.

[25] Q. Wu and D. X. Zhou, Analysis of support vector machine classification, submitted, 2004.

[26] Q. Wu and D. X. Zhou, Support vector machine classifiers: linear programming versus quadratic programming, Neural Comp. **17** (2005), in press.

[27] Y. Ying and D. X. Zhou, Learnability of Gaussians with flexible variances, submitted, 2004.

[28] T. Zhang, Leave-one-out bounds for kernel methods, Neural Comp. **15** (2003), 1397–1437.

[29] D. X. Zhou, The covering number in learning theory, J. Complexity **18** (2002), 739–767.

[30] D. X. Zhou, Capacity of reproducing kernel spaces in learning theory, IEEE Trans. Inform. Theory **49** (2003), 1743–1752.

[31] D. X. Zhou, Density problem and approximation error in learning theory, preprint 2003.

[32] D. X. Zhou and K. Jetter, Approximation with polynomial kernels and SVM classifiers, Adv. Comput. Math., in press.